The evolutionary genomics of CTCF binding and functional signatures in mouse

Dhoyazan Mohammed Ali Azazi

EMBL – European Bioinformatics Institute

Darwin College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

February, 2020

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

This document does not exceed the word limit of 60,000 words as defined by the Biology Degree Committee.

Dhoyazan Mohammed Ali Azazi
February, 2020

The evolutionary genomics of CTCF binding and functional signatures in mouse

Dhoyazan Mohammed Ali Azazi

Summary

Genetic differences within and between species predominantly lie in the noncoding sequence of the regulatory regions of the genome whose function and significance largely remain poorly understood. Despite significant progress in the field of genomics and the rapid progress in sequencing methods and the subsequent explosion of genomic data, our understanding of the role of the non-coding genetic sequence in the regulation of tissue- and species-specific gene expression is still lagging behind, limiting our comprehension of the evolutionary mechanisms and pressures that shape those expression profiles, and their involvement in the health and disease.

The CTCF protein demarcates mammalian genomes into discrete transcriptionally active domains, providing the platform for complex spatial and temporal regulatory processing of genetic information that govern biological processes. In this thesis, I investigate the dynamics and functional implications of evolutionarily novel CTCF binding sites in two *Mus* genus mouse subspecies, *Mus musculus domesticus* and *Mus musculus castaneus,* separated by a short evolutionary time of only one million years. The project investigated the subspecies-specific binding of CTCF in terms of the repeat content, evolution, functional impact and involvement in chromatin conformation. The key findings of this investigation are: (1) the incorporation of young CTCF sites into the non-coding genome via action of transposable elements is followed rapidly with the exhibition of various characteristics of biological function; (2) Unlike other tissue-specific transcription factors, allele-specific CTCF occupancy is affected by *cis*- and *trans*-acting regulatory mechanisms that exhibit similar functional characteristics; (3) CTCF evolutionary dynamics support both maintenance of pre-existing structures and functions and provide template for novel ones.

In summary, this thesis discusses the evolutionary dynamics of CTCF genomic occupancy and functional signatures in short evolutionary time, and

illustrates how either novel species-specific CTCF sites, or common sites with newly-acquired genotypic variants integrate into existing genomic architecture and begin to exert their effects.

To Mahasin

Acknowledgments

First and foremost, this thesis would not have been possible without the guidance, vision, support and encouragement of my PhD supervisor, Dr. Paul Flicek. The belief he has had, and instilled, in me and my independence has made this work a very rewarding experience. This thesis would not have been the same without the critical insight, constant feedback and excellent eye-for-detail provided by Dr. Maša Roller throughout the writing of this thesis. Their help has been truly instrumental in its success, and for that I am greatly indebted.

I would like to extend my thanks to all members, current and former, of the Flicek Research Group in no particular order: David Thybert, Camille Berthelot, Emily Wong, David Martin-Galvez, Thomas Rensch, Vasavi Sundaram, Elsa Kentepozidou, Petra Korlevic, Ericca Stamper and Martina Rimoldi. The rich and stimulating environment created through presentations, group meetings and discussion was a reliable source of invaluable feedback.

All the work carried out in this thesis is the result of the indispensable experimental efforts of our collaborators in the laboratory of Duncan Odom of the Cancer Research UK Cambridge Institute. It has been a pleasure working with Christine Feig who has been my main experimental collaborator, and the advice and input of Dr. Odom has been crucial to this work.

I would like to also thank the members of my Thesis Advisory Committee: John Marioni, Kyung-Min Noh and Jason Carroll, who

provided feedback, commentary and objective critique of my progress have always been a source of fresh and insightful perspective. A special mention goes to my former supervisors during my undergraduate and master research projects, Neil Bradman and David Balding, who provided the much-needed mentorship to steer me in the direction of research.

I am eternally in gratitude to the Darwin Trust of Edinburgh for their generosity in sustaining me throughout my doctoral work via the Jeff Schell Darwin Trust of Edinburgh studentship. Thanks are additionally due to the EMBL PhD programme for financial and moral support throughout the years of my PhD studentship. I would like to extend thanks to the EMBL Graduate and the EBI Research offices personnel, former and current, for their eagerness to help whenever the need arises.

I have been proud to be a Darwin College student at the University of Cambridge. The great science and the unparalleled atmosphere this city creates had to be lived to be believed.

I would not have been the person I am today without my parents, Fathia Noman and Mohammed Azazi. My parents fostered in me, as early as I can remember, a love of science and the ambition to follow on my dreams wherever the road takes me. A very special thanks to my brother, Alaa Azazi, who along with fraternal love and support, provided technical help and insight whenever I needed him most, even when the time difference between Cambridge and Calgary meant he went to sleep at dawn. I would also like to thank my sister whose love and trust in my abilities have always been a source of assurance and sustenance.

None of this would have been achievable without the unwavering love, never flinching passion, firm belief and unrelenting support of my beloved wife, Mahasin Abdu-Allah. She stood by me through the hardest of it all, and made it possible for me to devote my time and effort to getting where I am now, and for that I will be eternally be yours.

Contents

List of Figures

List of Abbreviations

3C	chromosome conformation capture
3D	three-dimensional
4C	circular 3C
5'UTR	5' untranslated region
APP	amyloid precursor protein
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	adenosine triphosphate
BIC	Bayesian information criteria
bp	base-pair
BS-seq	bisulfite treatment with next-generation shotgun sequencing
cDNA	complementary DNA
CEBPA	CCAAT/enhancer-binding protein alpha
CGI	CpG island
ChIA-PET	chromatin interaction analysis by paired-end tag sequencing
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CT	chromosome territory
CTCF	CCCTC-binding factor
DHS	DNase hypersensitive site
DNA	deoxyribonucleic acid
ENCODE	Encyclopaedia Of DNA Elements
eQTL	expression quantitative trait loci
eRNA	enhancer RNA
ERV	endogenous retrovirus
ESC	embryonic stem cell

FISH	fluorescent in situ hybridisation
FOXA1	forkhead box protein A1
GTEx	Genotype-Tissue Expression
GTF	general transcription factor
GWAS	genome-wide association study
HCP	high CpG content promoter
Hi-C	high-throughput sequencing chromosome conformation capture
HNF4A	hepatocyte nuclear factor 4 alpha
INDEL	insertion-deletion
INEs	Interspersed Elements
IP	Immunoprecipitation
LAD	lamina-associated domain
LCP	low CpG content promoter
LINE	long-autonomous interspersed element
lncRNA	long ncRNA
LTR	long terminal repeat
mESC	Mouse ESC
MLL3/4	myeloid/lymphoid or mixed-lineage leukemia protein 3/4
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	next-generation sequencing
PARP1	poly [ADP-ribose] polymerase 1
PCR	polymerase chain reaction
PIC	pre-initiation complex
PRC	polycomb repressive complex
PWM	position weight matrix
RNA-seq	RNA sequencing
RNA	ribonucleic acid
RNAi	RNA interference

RNAP	RNA polymerase
rRNA	ribosomal RNA
SBS	sequencing by synthesis
scRNA-seq	single-cell RNA-seq
SINE	short-autonomous interspersed element
SMC	structural maintenance of chromosomes
SNP	single-nucleotide polymorphism
SNV	single-nucleotide variant
TAD	topologically associating domain
TATA box	TA-rich sequence
TBP	TATA-binding protein
TE	transposable element
TF	transcription factor
TFBS	transcription factor binding site
TPRT	target-primed reverse transcription
tRNA	transfer RNA
TSS	transcription start site

Chapter 1

Introduction

The transformation of a few pluripotent cells into fully-formed multicellular organisms stems from the differentiation of those cells into the various specialised types and cell-lines. This is fundamentally governed by the complex array of regulatory networks that govern how the original set of identical DNA molecules in those progenitor cells are expressed in response to the innate developmental program and environmental cues. Understanding how the genome translates those biological and chemical signals into the spectrum of gene expression seen in the multitude of cell-types is a key question in the field of genomics.

The DNA-encoded information in eukaryotic genomes is first transcribed into a messenger RNA (mRNA) by RNA polymerases. RNA polymerase II (RNAP II) is responsible for the transcription of protein-coding genes. However, a large body of evidence points to the non-protein-coding regions of the genome as the location for the vast majority of genetic variants that influence the inter-individual and inter-species differences in phenotypic traits[1, 2]. Almost 90% of single nucleotide polymorphisms (SNPs) found to be associated with complex diseases are in the noncoding genome (40% in introns and 40% in intergenic regions) as revealed by the meta-analysis of 151 genome-wide association studies (GWAS) [3-5]. Consortium efforts in the last decade have been successful in revealing millions of DNA noncoding elements with putative regulatory potential across >100 human cell lines that could potentially explain the myriad of biological phenomenon in health and disease[6-9]. Whereas approximately 70% of protein coding sequences are evolutionary conserved, of the regulatory elements identified in the pilot ENCODE project only 10% were evolutionary constrained[6, 10].

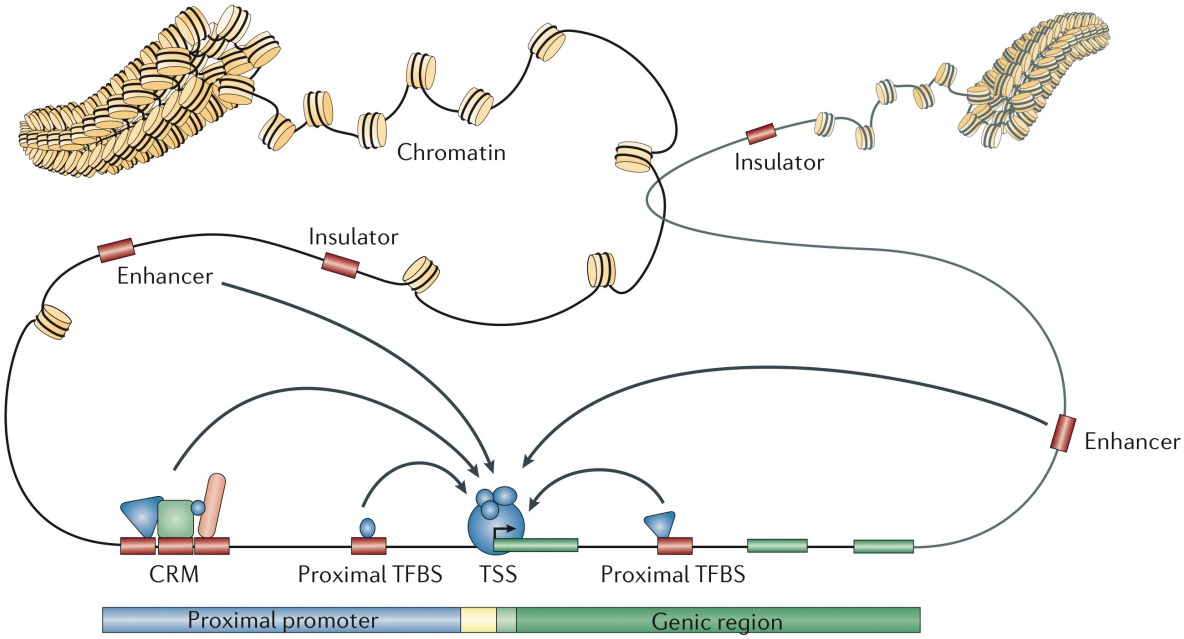


Figure 1.1: Regulation of gene transcription

A summary view of the principle components of the transcriptional machinery. DNA is wrapped around histones in the chromatin, forming nucleosomes. Boundaries are demarcated by insulators. The transcription start site (TSS) comprises the proximal- and core-promotors, and binds RNA-polymerase II (blue circle) with its associated general transcription factors (little blue circles on top). Transcription factors (TFs, blue triangles/circles, pink sticks and green squares) bind to transcription factor binding sites (TFBSs) near the TSS (proximal elements) or far away (enhancers). TFBSs also form *cis*-regulatory modules (CRMs) when bound in clusters. Arrow indicate the various interactions acting on the gene promotor. Figure adapted from Lenhard *et al.*[11].

Gene expression is critically regulated at gene level to ensure the homeostasis of the cell and fidelity of the developmental programme of the organism. The control of this regulation is shaped by the interactions of core transcriptional machinery, the communication with particular transcription factors and the three-dimensional (3D) conformation of the genome[12]. This is made possible by the presence of discrete *cis*-regulatory DNA elements, known as enhancers, that control the spatiotemporal pattern of gene expression in tissue- and species-specific manners in the eukaryotic genomes[13]. This in turn is achieved through the action of multiple proteins/protein complexes that facilitate the interaction between these elements and the DNA at coding gene[14].

Transcriptional enhancers are integral to the regulatory processes in the cell as they integrate both intrinsic inputs and extracellular signals to recruit transcriptional activators or repressors[15, 16]. The folding of chromosomes into favourable 3D structures allows for further complex patterns of transcriptional control by establishing physical links between genes, their regulatory elements, and the surrounding chromatin[17]. In the last few years, advances in the understanding of the nature of non-coding regulatory elements have allowed for the genome-wide mapping of transcription factor binding sites, long-range DNA interactions and chromatin signatures using assays based on chromatin accessibility and conformation capture [7, 18-24].

In this thesis, I have investigated how a master regulator of the genome, the CTCF protein, known to be involved in the various levels of transcriptional regulation has evolved to exert those functions in short evolutionary time scales. This investigation was based on a computational approach, focusing on the binding patterns of CTCF and the functional signatures of its genomic occupancy in tissue- and species-specific manners. The evolutionary genomics governing the inheritance of CTCF were explored in an attempt to understand how evolutionarily variant CTCF sites are inherited and what patterns are observed by their binding in species-specific and hybrid biological contexts. This work was facilitated by the technical breakthroughs in next-generation sequencing (NGS) techniques, and the availability of high-quality genome sequences for the species involved in those investigations[25].

Therefore, in this introductory chapter, I will give an overview of the field of transcriptional regulation and evolutionary genomics, as well as the methods relevant to the work conducted in this thesis. This chapter starts with a broad look at the levels of regulation of gene expression from promoter to sequence variants, with particular emphasis on the roles CTCF plays in this process. It then moves into an overview of the field of evolutionary genomics, with particular emphasis on the evolution of transcriptional regulation. This chapter concludes with a brief description of the methods and approaches employed in the investigations performed and reported in this thesis.

1.1 Regulation of gene expression

Regulation of gene expression is crucial to the development and maintenance of cellular processes of adult multicellular organisms. Transcription is the first and foremost step in this heavily regulated process and is where most of the rate-limiting steps are found.

Different cell types vary in their mRNA content and this variation correlates with the protein products of cell-type specific genes[26]. Failure to regulate transcription results in many disorders, particularly cancers[27, 28]. Regulation of transcription is achieved through the compound action of DNA *cis*-regulatory elements: core-promoters, promoter-proximal[11] and -distal elements such as enhancers[13, 29], repressors[30, 31], insulators[32] and boundary elements[33](Figure 1.1).

Transcription initiates at the 5' end of genes, within elements known as core promoters. Core promoters are short sequences of about 100 base pairs (bp) that surround the transcription start sites (TSSs) that recruit RNAP II, along with the rest of transcription factors required for the formation of the pre-initiation complex (PIC). This process determines the proper positioning and orientation for the initiation of transcription[34]. Core promoters, nonetheless, cannot maintain transcription, and on their own they yield basal transcriptional activity. Cell-type specific expression is further modulated by the integration of proximal and distal elements, enhancers[16, 35]. Enhancers are DNA sequences, a few hundred bp in length, that function as platforms to recruit transcription factors to bind to specific DNA sequence motifs (reviewed in Spitz and Furlong[35]). Enhancers are capable of activating transcription regardless of their location, distance and direction to gene promoters. There are enhancers that were even observed to promote the transcription of olfactory receptor genes present on different chromosomes[36, 37]. It remains a challenge in biology to connect the regulatory elements to the genes they control since most of these are often separated by thousands of bps [38, 39].

In addition to *cis*-regulatory elements, the unfolding of chromatin is a prerequisite to the activation of transcription. The decompaction of chromatin is facilitated by the action of enhancer-bound transcription factors that attract histone-modifying enzymes, or through an ATP-dependent mechanism (more on this later in 1.1.3)[40]. This process increases DNA accessibility for the various components of transcriptional machinery to assemble and initiate mRNA production.

The following sections go over the main layers of transcriptional regulation of gene expression: (1) regulation at the level of transcription initiation at the core promoter; (2) regulation of *cis*-regulatory elements via binding to transcription factors; (3) regulation at the level of chromatin through nucleosome modifications, looping and higher order structures and chromatin compartmentalisation; and (4) regulation via *cis*- and *trans*-acting variation.

1.1.1 Transcriptional regulation at the core promoter, initiation and PIC formation

The classical definition of a gene promoter is that it is the DNA region required for the initiation of gene transcription[41]. Based on this definition, these regions overlap with the TSSs, the loci where the output of regulatory potential translates into gene expression via initiation of transcription. The fundamental function of the promoter is to provide a binding site for the assembly and positioning of the pre-initiation complex (PIC), which in turn recruits a DNA-dependent RNA polymerase (such as RNAP II). In eukaryotes, this region is termed the "core promoter"[11]. The core promoter consists of multiple interchangeable sequence elements, such as an initiator element and a TA-rich sequence (TATA box), that bind the component of the PIC[42]. These elements are needed for the formation of the basal transcriptional machinery[43].

Whereas bacterial and archaeal gene transcription is carried out by only one RNA polymerase (RNAP), eukaryotic genomes are transcribed by three RNAPs that are highly conserved in evolution[44]. RNAP I is responsible for the synthesis of the ribosomal RNAs (rRNAs) RNA precursors of the translational machinery. RNAP II, as discussed above, transcribes all protein coding genes to mRNA. RNAP III is involved in the transcription of transfer RNAs (tRNAs) and small subunit of rRNA, in addition to small untranslated RNA such as general transcription factors (GTFs). Whilst the three RNAPs share a common catalytic core, essential polymerase-specific subunits and GTFs are needed for the proper regulation of each polymerase's activity[45]. Importantly, the basal transcriptional activity of RNAP II is a concerted effort of the RNAP II complex with its associated GTFs: TFIIA, TFIIB, TFIIE, TFIIF, TFIIH and TATA-box binding protein (TBP). These components assemble together to form the PIC in a step-wise process[46, 47]. Additionally, during promoter melting, an ATP-dependent translocase activity of the GTF TFIIH is required by RNAP II in order for the transcriptional bubble to form[48-51].

The process of DNA transcription contains at least eight (Figure 1.2) principal steps that can be regulated in a rate-limiting manner to ensure the precise control of gene expression. This is better understood as a cycle that starts with (1) RNAP II accessing the obstruction-free core promoter after being cleared from nucleosomes. (2) RNAP II and associated GTFs form the PIC on the core promoter. (3) Promoter opening follows, and transcription is initiated. (4) RNAP II escapes from the core promoter and begins early elongation of the nascent RNA chain, and moves into the promoter-proximal pause region. (5) Paused RNAP II is subsequently hyperphosphorylated, and clears the pause region, resulting in either the termination of transcription or continuing elongation of the mRNA chain. (6) If RNAP II undergoes

continuing elongation, it proceeds throughout the whole length of the genebody. (7) Transcription is terminated once RNAP II reaches the end of the gene, and (8) a new cycle of transcription is initiated if the conditions at (1) remain permitting[12](Figure 1.2). The distribution of RNAP II across many genes have been studied in a number of species as a proxy to the rate-limiting steps in transcription[52]. A wealth of data on RNAP II density genome-wide is now available for multiple species such as yeast[53], fruit fly[54, 55] and humans[56].

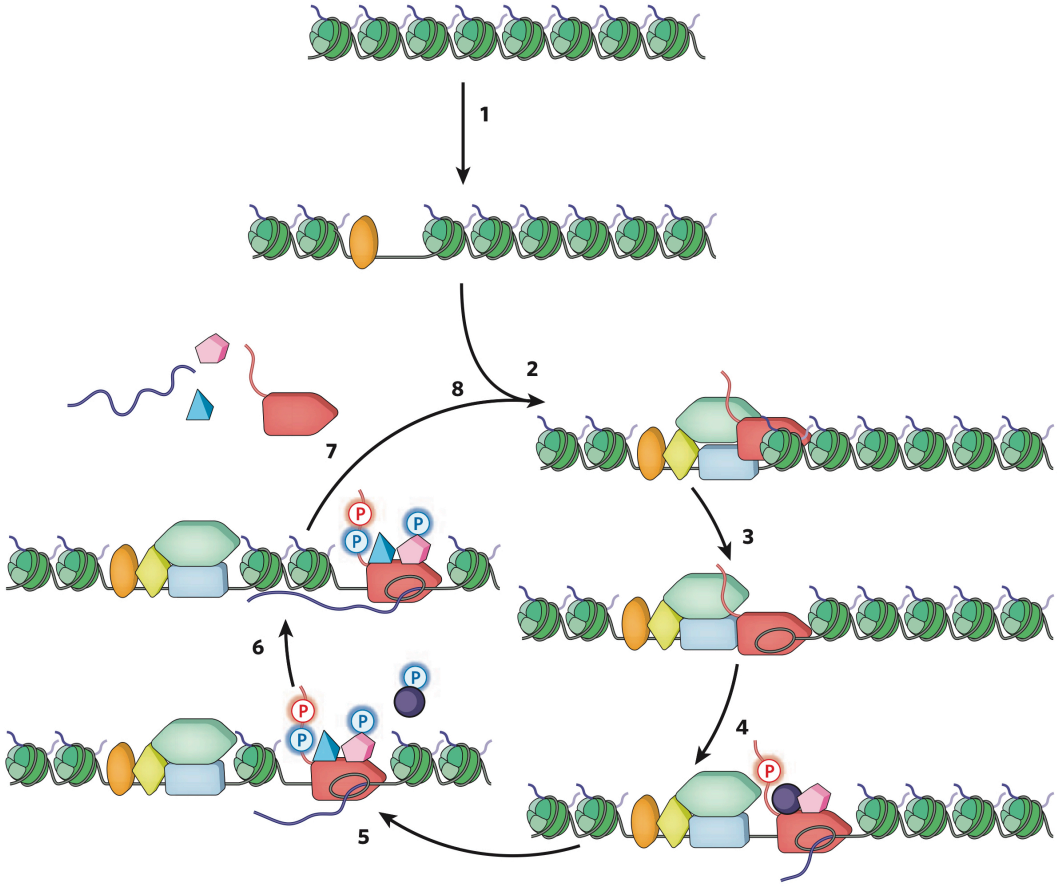


Figure 1.2: The 8-step transcriptional cycle

(1) DNA is packaged into nucleosomes (green), until an activator (orange oval) binds and nucleosome remodelling begins. (2) A 2nd activator (yellow diamond) promotes GTFs (blue rectangles) binding and attract coactivators (green hexagon). RNAPII enters the PIC. (3) DNA is unwound (oval inside RNAPII). (4) RNAPII begins transcribing 20-50 bp (purple line), then pauses, mediated by SPT4—SPT5 (pink pentagon) and negative elongation factor (purple circle). Ser 5 residues of RNAPII C-terminal domain (CTD) are phosphorylated (red P). Step 5: P-TEFb (blue triangle) phosphorylates Ser 2 of RNAPII CTD, SPT5 and the NELF subunits (blue Ps). NELF

dissociates from the complex. RNAPII exits the pause. Step 6: Nucleosomes are disassembled and reassembled as RNAPII elongation complex proceeds along the gene. Step 7: After RNAPII complex transcribes the gene, it is removed from the DNA, and the mRNA is released. Step 8: The freed RNAPII reinitiate. Figure adapted from Fuda *et al.* [12]

Regulation of transcription at the first step, for example, is established by limiting RNAP II access to core promoters that are covered by nucleosomes. Transcription of such genes requires the recruitment of nucleosome-remodellers and -modifying enzymes. This is achieved by communication between the core promoter and proximal regions and more distal enhancer sequences, bound by activator or repressor transcription factors, indirectly via a Mediator complex of proteins[57]. Human and yeast activators have been shown to interact with the SWI/SNF remodelling complexes to allow nucleosome clearing and activation of transcription[58]. The recruitment of histone acetyltransferase Gcn5 to galactose-inducible genes by Gal4 in yeast modifies the chromatin state at the promoter and allow for transcription initiation[59]. RNAP II recruitment provides another step for regulation of transcription. Following initiation, RNAP II proceeds rapidly towards elongation and becomes uniformly distributed across the gene body[60]. In response, PIC formation is accelerated by activators that interact with the GTFs TBP, TFIIA, TFIIB and TFIID[61]. The recruitment of the Mediator complex leads to further interaction with the GTFs, stabilization of the PIC, increased RNAP II recruitment and upregulated expression[62].

Another layer of regulation at the core promoter is based on DNA sequence content around the promoter-proximal regions in the genome. Stretches of ~ 1000 bp long with a base composition rich in GC-bases and high density of CpG dinucleotides are known as CpG islands (CGIs), and are prominent in mammalian genomes[63]. CGIs account for 1% of the genome, and incorporate the TSSs of most mammalian genes. They are significantly more enriched in GC-content that the bulk genomic DNA in humans (65% vs 40% G+C content)[64]. CGIs are found in the TSSs of about 60% of genes in human, and they can also be found within and between genes. Non-TSS CGIs additionally exhibit TSS-like characteristics, such as association with RNAP II, detection of present transcripts and colocalization with the trimethylation of lysine 4 of histone H3 (H3K4me3) chromatin mark of transcriptionally active promoters, suggesting their regulatory potential as promoters[65-67].

Based on their CGIs content, mammalian promoters can be classified into three major types: Type I (adult), Type II (ubiquitous) and Type III (developmentally regulated)[11]. Type I promoters are responsible for the tissue-specific gene expression

in terminally differentiated adult cell lines. They are characterized by a narrow/sharp TSS and disordered nucleosome conformation, consistent with their spatiotemporal narrow range of expression. These promoters are also, crucially, the ones whose promoters are low in CGIs[68-71]. Type II promoters are expressed broadly throughout the organism, and are defined by a broad TSS and an ordered nucleosome structure. They are also the promoters highly enriched for CGIs[68-71]. Type III promoters are variably regulated during development and exhibit Polycomb-repression with a broad histone methylation marks (H3K27me3), and CGIs that extend into the genebodies[72].

CGIs have additionally been shown to be nucleosome-deficient *in vivo*, making CGI-rich promoters easily accessible for transcription factors, and they do not require the ATP-dependent SWI/SNF remodelling of chromatin discussed earlier[73]. Experiments in mouse brain cell lines indicate that promoter-associated, non-methylated CGIs exert chromatin modifications by interaction with Cfp1 and other CGI-binding proteins[74]. Therefore, CGIs functional role in transcriptional regulation lies in its inherent ability to expose local chromatin of nucleosome, helping it adopt a transcriptional initiation-friendly configuration[64].

1.1.2 Transcriptional regulation at distal enhancer elements via TF binding

Enhancers are *cis*-regulatory elements located a considerable distance upstream or downstream from the core promoters that recruit transcription factors, RNAP II and chromatin remodellers to establish or maintain an active transcriptional state via PIC [75-80]. Enhancers interact with the core promoter via the Mediator complex and the GTF TFIID to help attract RNAP II to the PIC[81]. They modulate transcription of the genes they control, which may not be the closest ones based on their distance alone[82-88]. These interactions are not exclusively used for the initiation of transcription, but may also contribute to the release of RNAP II from the promoter-proximal pausing region (step 4 discussed above)[89]. The first mammalian enhancer was identified downstream of the immunoglobulin (*Ig*) heavy-chain locus, which upregulates the expression of the *Ig* gene in the lymphocyte-derived cells during B-lymphocyte differentiation[90, 91].

Enhancers begin recruiting general and lineage-specific transcription factors at the ESC stage, whereas promoters are unlikely to be bound by developmentally important and tissue-specific factors[92-97]. Thus, enhancers play a central role in establishing the spatiotemporal pattern of gene expression in animals[13, 98-101].

Highly-expressed genes that share common regulatory enhancers tend to cluster together at nuclear loci called transcription factories[102]. Enhancers are also capable of activating homologous and heterologous promoters independently of their position up/downstream of the genes or their orientation, and additionally display DNase I hypersensitivity, which is the main criterion used to identify enhancers in genome-wide scans of mammalian genomes[43].

A new class of regulatory enhancers has been recently described as super-enhancers[103-105]. These are enhancer-like *cis*-regulatory elements marked by the acetylation of histone H3 lysine 27 (H3K427ac) and occupied by master regulators, especially the Mediator complex. They form a cluster of regulatory elements up to 12.5 kb in size, flanked by CTCF-binding sites, indicating that their influence may be regulated by boundary elements[106]. They have been identified in several cell lines, and were proposed to act as critical switches to determining cell fate[107-109]. However, a recent study using an erythroid cell line in a mouse model that targeted the constituents of the α-globin super-enhancer individually for deletion demonstrated that each element acts independently and in additive manner to exert their phenotype. No evidence of synergistic activity between the super-enhancer's members or higher-order effects were clearly observed[110].

In addition to recruitment of RNAP II to promoters, a number of studies in mammalian genomes have recently demonstrated that RNAP II is similarly recruited to active enhancer elements where is initiates widespread transcription of their DNA sequences[111-113]. Enhancer-transcribed RNAs (eRNAs) are 0.5-5 kb in length, similar to long-noncoding RNAs (lncRNAs)[114]. eRNAs had originally been considered as sequencing noise of nonspecific mRNA transcripts resulting from genomic regions accessible to transcriptional machinery[115], but there is a current agreement that these eRNAs may have a role to play in gene regulation, although the nature of this role remains controversial[43]. *In vivo* study in 2014 confirmed that many of these eRNAs are expressed in tissue-specific manner[116], followed a comprehensive transcriptomic profiling of eRNAs in humans[117]. A 2015 study demonstrated the role of the RNA exosome in regulating eRNA degradation[118], and another study showed some eRNAs are implicated in metabolic stress[119].

In order for enhancers to activate transcription, they require the binding of multiple TFs to ensure the integration of both cellular signals and extracellular cues from the environment[107, 120, 121]. The action of multiple TFs is critical to the function of enhancers as these elements have a high affinity to nucleosomes[122], hence the compound activity of all these factors provide a strong barrier to the repressive effect of chromatin on the underlying regulatory elements. Enhancer elements are thus

thought of as genomic 'nexus' sites where the input activities of myriad TFs are integrated into the overall array of transcriptional regulatory output. These TFs are assembled into a hierarchical logical organisation at the enhancer elements, as shown in a recent study using mouse liver *cis*-regulatory modules as a model for this interaction[123].

The interaction between the various TFs bound to enhancers in close proximity is thought to happen cooperatively and plays a central role in nucleosome eviction and subsequent enhancer activation[35]. Three main modes of cooperativity between TFs at enhancer elements have been proposed. The first mode, termed 'direct cooperativity', relies on the direct physical association between TFs prior to or concurrent with binding to their DNA motifs. The second mode, 'indirect cooperativity' or 'collaborative competition', occurs when a number of TFs compete over access to the enhancer element with the same histone octamer, as predicted *in silico* modelling[124], then experimentally demonstrated[125]. As a result, the more TF motifs found at an enhancer, the higher the rate of nucleosome eviction, DNA binding and gene expression; a view supported by multiple studies investigating endogenous enhancers and synthetic reporter assays[126, 127]. A third mode is observed at a number of developmental enhancers, where evidence of step-wise activation by lineage-determining master regulators or 'pioneer factors' that bind directly to nucleosomal DNA and prime enhancers for action[128]. Pioneer factors are capable of recruiting chromatin remodellers, histone-modifying enzymes, easing the burden of direct competition between subsequent TFs and coactivators and nucleosomes on binding to the DNA[129].

Although pioneer factors are supposed to have universal remodelling and targeting capabilities, they have been demonstrated to display a level of cell-type specificity[129-131], influenced by various factors such cell-type-specific cofactors (e.g. FoxA1 and Sox2)[132-134], signalling (TNFα)[135] and chromatin state[136, 137]. A comprehensive investigation of the genomic occupancy of three such factors, FOXA2, GATA4 and OCT4, in multiple human cell lines have shown that all three TFs display tissue-specific occupancy, even when their expression is imposed ectopically. The expression of additional cofactors increased the enrichment at a subset of binding sites in alternative cell types, indicating the essential requirement of such cofactors for the correct function of pioneer factors[138].

Coactivators are factors essential for the proper function of the DNA-binding TF, but not necessarily for the basal transcriptional machinery, and they do not exhibit site-specific binding on their own[81]. They exert their effects by modifying the chromatin context of enhancers using a host of histone-modifying enzymes: histone

acetyltransferases (p300/ CBP, SAGA complex, MOF, TIP60), histone methyltransferases (MLL3/4, CARM1), chromatin remodellers (Brg1, CHD7), and the Mediator complex that facilitates the interaction with the PIC at the core promoter [81, 139, 140]. Therefore, coactivators are linked to most active enhancers independently of the cell-line they are found in, allowing their use for annotations of putative enhancers in a variety of tissues, using ChIP-seq to map their genomic occupancy [141-143].

Enhancers adopt different, but not mutually exclusive configurations to interact with their corresponding promoter based on the genomic distance between them. Three principal models have been proposed: a linking model, a tracking model, and a looping model[43]. In the linking model, a cohort of TFs are sequentially recruited after the binding of the first activator or pioneer TF which results in open chromatin conformation. A trail of TFs subsequently bind along the chromatin fibre extending from the enhancer element towards the core promoter, attracting the PIC for transcription initiation[134]. This model, however, is only applicable at short ranges as the formation of such a cascade is highly unlikely over vast genomic distances. In the tracking model, TFs occupying the enhancer element include active RNAP II and proceed towards the target promoter unidirectionally[144]. The most well-known example of this model is the 70-kb region containing the locus control region (LCR) of the human beta-globin gene[145].

The most common model, however, involves the direct contact between promoters and enhancers by looping of the chromatin to bring close regions separated by long distances (Figure 1.3). The resultant loops are further stabilised by protein-protein interactions between the TFs, coactivators, GTFs and RNAP II and the PIC. Specialised factors and protein complexes work to bridge the gap and allow physical contact between the *cis*-elements they occupy. These factors include the chromatin remodelling Mediator complex and the CCCTC-binding factor (CTCF)-cohesin protein complex[146] (Figure 1.3). In addition, eRNAs have being suggested to contribute to the process by stabilising enhancer-promoter looping via interaction with either the cohesin or Mediator complexes[147, 148].

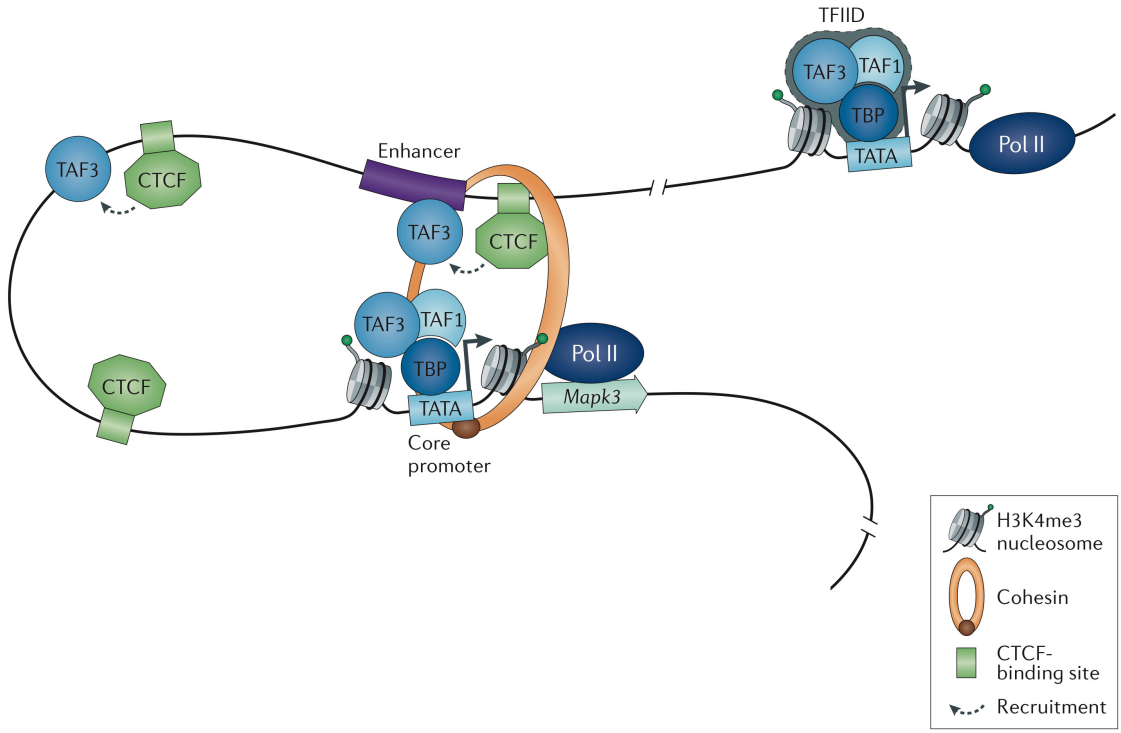


Figure 1.3: Enhancer-promoter interactions

Enhancer-promoter interactions allow the tethering of distal *cis*-regulatory elements to the core-promoter. Recruitment of TATA-binding protein-associated factor 3 (TAF3) at endodermal enhancers by the CTCF-cohesin complex and chromatin looping activates the mitogen-activated protein kinase 3 (Mapk3) gene in mouse embryonic stem cells (mESCs). Figure adapted from Ong and Corces [80].

Functional enhancers harbour clusters of TF recognition motifs. These are DNA sequences that the DNA-binding domain on TFs can recognize and occupy in order to exert their effects. The discovery and annotation of these motifs has been the subject of large-scale experimental and *in silico* efforts leading to the establishment of two major databases of such information: JASPAR[149] and TRANSFAC[150]. The annotations included are based on experimentally determined cohorts of TF binding sites (TFBSs) that are subsequently aligned and the enrichment of bases at each position of the motif weighted to produce positional weight matrices (PWMs) for the plethora of TFBS in the eukaryotic genome[151].

The identification of genome-wide enhancer elements and TF motifs have opened the door for recent investigations attempting to link genetic variation in the noncoding genomes with particular phenotypes in health and disease. A study in 2013 estimated that 7.5% of variation in TF-DNA binding events can at least be explained

by alteration in TFBS motifs[152]. Using genomic data from the Genotype-Tissue Expression Project (GTEx)[9], a recent group showed that purifying selection in the general population has reduced haplotypes predicted to increase pathogenic coding variant penetrance, indicating that *cis*-regulatory variation could predispose to disease risk by modulating the penetrance of gene variants [153]. Another study used a CRISPR-Cas9 system to target DNase I hypersensitive sites of the beta-globin and HER2 loci in humans, affirming the role of known regulatory regions, and discovering previously unreported elements[154]. A similar method probing the function of enhancer clusters in mouse embryonic stem-cell lines (ECs) revealed the effect of deleting these enhancers varies greatly, with reduction in gene expression ranging from 12% to 92%[155]. Partial deletions of one or more components of these clusters also demonstrated a degree of redundancy in gene regulation by these enhancers, a feature that may ensure the fine-tuning of transcriptional output and protect against arbitrary loss of some of these enhancers[155]. Views from cancer research support the importance of enhancer sequence fidelity for the proper functioning of these elements, with evidence from the analysis of 102 tumour cell genomes that somatic small insertion or deletions (INDELs) can nucleate oncogenic enhancer activity[156, 157]. An insertion in the enhancer of the LMO2 oncogene leads to its activation and the progression to leukaemia[157].

It is still a long road towards understanding how the interactions between the regulatory elements and our ability to predict the outcome of such interaction on the genomic-scale. The target genes for most enhancer elements are unknown, and whilst it is well-agreed that many genes are under the control of several regulatory elements[158, 159], our models of the multi-modular structure of gene regulation are still limited.

1.1.3 Transcriptional regulation via modifications to the chromatin structure

The 3D conformation of the genome performs a major role in the maintenance of genome stability, organization, and regulation of gene transcription in health and diseases. Chromatin is the most intricately regulated ensemble in the cell. It is made up of genomic nuclear DNA and its associated proteins and RNA molecules. DNA-associated proteins include histones, TFs, the PIC, mRNA, coactivators and other complexes necessary for the replication and repair of the 3D genome. Although chromatin folding is governed by a set of general principles in all cells of the multicellular organism, the spatial configuration of the genome is highly variable

between cell lines to the extent that no two nuclei will exhibit the same range of chromosomal connections[160-162].

A major determinant of DNA accessibility is the presence of nucleosomes. A nucleosome-wrapped DNA sequence is more resistant to the binding of TFs that the same sequence in a nucleosome-free state[163, 164]. A nucleosome is evicted from the DNA by action of remodelling factors that alter its location along the DNA molecule[40]. Nucleosome-depleted regions provide an amble opportunity for DNase I activity. There are over 870,000 DNase I hypersensitivity regions in humans, covering almost 9% of the genome[10]. These areas of open chromatin tend to cluster around the TSSs and coincide with CTCF binding sites when identified in multiple cell lines. However, cell-type-specific sites are found further away from the TSSs and harbour motifs recognized by the corresponding tissue-specific TFs[10].

Regulation of transcription at the level of chromatin also involves the post-translation modification of the histones that make up the nucleosome complex[165].An extra layer of regulation of the chromatin is conferred by the Polycomb and Trithorax protein complexes[166-168]. Chromosomes are additionally partitioned into nuclear territories and compartments, made up in turn of topologically-associated domains (TADs), each involving a number of entangled DNA loops, formed by action of several proteins such as CTCF and the cohesin protein complex [169, 170]. Major development in chromosome conformation capture technologies in the last decade, coupled with advancements in mathematical modelling for interpreting the data, has revolutionised the analysis of chromatin and elucidated its involvement in various cellular process[171-176].

The following section will give an overview of the three main levels of regulation of gene expression via changes to the chromatin structure: (1) Nucleosomes and histone modifications at *cis*-regulatory elements; (2) DNA-looping and higher-order chromatin conformations; and (3) chromatin compartmentalisation and the establishment of TADs.

1.1.3.1 Nucleosomes and histone modifications

In eukaryotes, nuclear DNA exists in complex with a particular class of proteins known as histones. A DNA sequence of 147 bp wraps around the nucleosome core particle in 1.7 supercoiled turns. The core nucleosome is comprised of two histones 3-histone 4 (H3-H4) and two histone 2A-histone 2B (H2A-H2B) dimers. A 10-80 bp DNA linker associated with the linker histone 1 (H1) separates nucleosomes, by promoting compaction of neighbouring nucleosomes. H1 is common in heterochromatin where

greater compaction results in its condensed appearance. This complex of DNA and nucleosome folds into a 10 nm diameter fibre. *In vitro* studies have shown that *in vivo* this fibre forms a helical fibre containing 6-11 nucleosomes per turn. This in turn folds further to make higher order chromatin fibres in interphase, and a 200-300 nm structures during condensation of mitotic chromosomes (reviewed in Felsenfeld *et. al*[177]).

Histone modifications lead to changes in nucleosome occupancy and regulatory potential. These are post-translation modifications that have been reported to affect over 60 different amino acid residues on histones, instigated by protein-modifying enzymes, many of which also have non-histone substrates[178]. These include: acetylation, methylation, phosphorylation, ubiquitylation, propionylation, butyrylation, formylation among others[179, 180]. Enhancer-bound TFs have recently been shown to actively recruit histone-modifying enzymes through direct interactions with histone tails, and ATP-dependent remodellers of chromatin that disrupt nucleosome-DNA contacts and allow nucleosome displacement along the DNA, and its removal or exchange[40, 89, 181, 182]. Histone modifications are critical for regulation of transcription. For example, the acetylation of lysine residues on histone tails neutralises their positive charge and changes chromatin overall charge[183, 184]. Similarly, lysine methylation may lead to an increase in the binding affinity on the DNA-binding domains on a number of factors believed to act upon chromatin packaging[185]. Disruptions to the histone modification process have been associated with a number of disease phenotypes[186-188]. This is not unlikely given their function as transducers of intrinsic signals from the cell to the genome[189].

The 'histone code' refers to the combinations of modifications required to instigate downstream events[165, 190] (Figure 1.4). It is now commonplace to refer to some of the modifications as 'activator' or 'repressor' marks depending on the outcome of the event they regulate[191]. There are currently over 150 described histone modifications, and a single nucleosome could carry multiple modifications at the same time, alas only a small number of these modification patterns have been discovered [67, 192]. Several groups have been working on profiling and mapping different histone marks genome-wide to identify the underlying code and how it associates with gene activation/repression, or other genomic features such as promoters, enhancers and insulators[21, 56, 193-195].

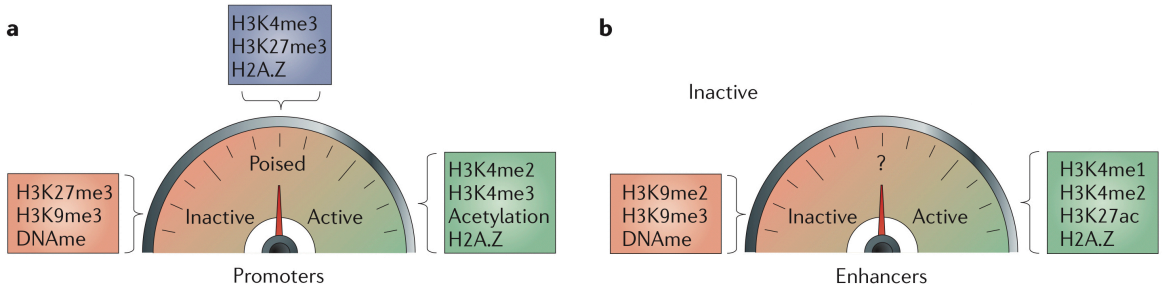


Figure 1.4: Histone modifications "code" for *cis*-regulatory elements

Various histone modification combinations dictate the regulatory potential of promoters (a) and enhancers (b). Figure adapted from Zhou *et al.* [196].

As discussed earlier in 1.1.1, most promoters colocalise with regions of high GC content, CGIs, and are known as high GC promoters (HCPs), in contrast to their counterparts in low-CGI regions, the low GC promoters (LCPs). Studies have shown that the histone mark H3K4me3 coincided with HCPs, and were characterised with increased chromatin accessibility, histone acetylation, binding of histone H3.3 and marked DNase I hypersensitivity[197-199] (Figure 1.4a). Similar to the HCPs, these accessible H3K4me3-marked regions were also hypomethylated at the DNA level[200]. On the contrary, LCPs appeared inactive by default, and they lacked any measurable enrichment with either H3K4me3 or H3K4me2 in ESCs or adult cell lines. However, a minor subset of LCPs bore the H3K4me3 mark and were highly expressed compared to the unmarked LCPs[21, 200] (Figure 1.4a).

Repressed promoters display a unique histone modification pattern that reflect their transcriptionally inactive state. They are usually marked by the tri-methylation of lysine 27 of its histone 3 (H3K27me3), which is also the prototypical mark of the Polycomb repressor complex. Polycomb repressor complexes, PRC1 and PRC2, inhibit transcription to maintain cell-type-specific gene expression patterns[201]. A large number of HCPs are targeted by Polycomb in mammalian genomes. For example, about 20% of HCPs in ESCs are bound by PRC2 and marked with its associated mark, H3K27me3[202-204]. Interestingly, these promoters also carry the H3K4me3 activator mark , thus are capable of 'bivalent' characteristics, being both activated and repressed[205]. ESC bivalent promoters show very low levels of gene expression[206], but later studies have identified some RNAP II enrichment[56, 207]. The tri-methylation of lysine 9 of histone 3 (H3K9me3) is another repressed promoter mark that correlates with constitutive heterochromatin and DNA hypermethylation[196].

Whereas mapping promoters is somewhat straightforward, histone modifications have been instrumental in helping identify enhancer elements in an

unbiased fashion[196]. Enhancers are characterised by the presence of particular histone marks in addition to the binding of TFs and other co-activators such as p300[75]. Analysis have shown enhancers to be enriched for marks such as H3K27ac, H3K4me2, H3K9me1, H3K27me1, H2BK5me1 and H3K36me1, indicating a degree of redundancy in the histone code[199] (Figure 1.4b). Nevertheless, an enhancer histone code could also be fine-tuned by acetylation of H2A.Z, resulting in corresponding differences in downstream gene activation[198]. Although enhancers are commonly marked with H3K4me1 and H3K27ac, they could also be marked with H3K4m3, the active promoter mark, if an enhancer is highly transcribed[195, 208, 209]. Therefore, putative active enhancers are identified by a cohort of criteria including measuring the ratio of H3K4me1 to H3K4me3, along with the presence of H3K27ac, the replacement of histones with the variant H2A.Z, the binding of coactivators p300/CBP and cooperative binding of master TFs[16, 210-215]. Poised enhancers, on the other hand, are characterised by the noted absence of the H3K27ac mark and the enrichment of H3K27me3 and/or H3K9me3, an epigenetic feature later found to be common a large number of enhancers with tissue specificity. This poised state could, however, be readily reversed when the histone mark H3K27me3 is modified and replaced with H3K27ac[209].

1.1.3.2 Long-range interactions and chromatin-loops

Chromatin status provides a proxy to the level of *cis*-regulatory activity, and a measure of the widespread changes in enhancer position and activation state in relation to the gene promoters they interact with. In addition to *trans*-acting TFs, activation of enhancers is combined with the folding of chromatin into loops of long-range interactions between *cis*-regulatory regions and the core promoters of the genes they regulate[79, 216, 217].

Chromatin long-range loops form when pairs of loci have stronger interactions than any of the other loci in-between[218]. ~30% of loops involve promoters and enhancers, resulting in changes in gene expression and transcriptional activity[219, 220]. 66% of active promoters interact with their nearest enhancers, yet 30% bypass that enhancer, whereas 4% display preferential unidirectional interaction even if the enhancer in the other orientation is much closer[221]. In contrast to previous theories, 90% of these promoters involve at least one more distant enhancer, leading to complex interaction patterns[221]. A recent major study in 17 human primary hematopoietic cell types found 17,500 interactions between promoters and promoter-interacting regions (median = 4 interactions/promoter), 50% of those interactions were with one promoter and 10% with =>4 promoters, further complicating the landscape of gene regulation through higher-order chromatin conformations[222]. This redundancy could

be explained from an evolutionary perspective as a safety mechanism to counter pathological effects of enhancer disruption by mutation[222].

A 2012 study looking at the 1% of the human genome from the ENCODE pilot project showed that long-range interactions display asymmetrical preference for elements located 120 kb upstream of the TSS. These interactions were found to not be blocked by CTCF and cohesin occupancy, and that only 7% of the looping interactions are with the nearest promoter, emphasizing that proximity is not definite indicator of long-range interaction. Furthermore, promoter-enhancer looping interactions correlated significantly with gene expression and the presence of eRNAs[39].

Using Promoter Capture Hi-C, a recent study generated a high-resolution atlas of chromosomal interactions in human pluripotent and lineage-committed cells for ~22,000 promoters. They identified putative target genes for known and predicted enhancer elements, and revealed how gain and loss of promoter interactions changes the dynamics of *cis*-regulatory contacts upon lineage commitment[223]. The Mediator and cohesin protein complexes have been shown to be implicated in promoter-enhancer looping interactions[224]. Promoter-enhancer interactions also restrict divergent transcription of noncoding RNAs (ncRNAs) made by RNAP II from bidirectional promoters by adopting gene-loop configuration[225]. Interestingly, formation of promoter-enhancer loops is also associated with recruitment of the corepressors NCoR and HDACs, demonstrating that chromatin looping is coupled to activation of poised enhancers. MLL3/4-dependent H3K4me1 has been shown to orchestrate long-range promoter-enhancer interactions in mammalian cells. Yan *et al.* demonstrated that increased levels of chromatin interactions correlated with MLL3/4-dependent deposition of H3K4me1 at enhancers in differentiating mouse embryonic stem cells. H3K4me1 loss resulted in reduced levels of chromatin interactions and defective gene expression during differentiation[226]. H3K4me1 facilitates cohesin complex recruitment to chromatin *in vitro/vivo*, potentially allowing MLL3/4 to mediate chromatin interactions between enhancers and promoters.

1.1.3.3 Topologically-associated domains (TADs) and chromatin compartmentalisation

The 3D configuration chromatin assumes is central to its ability to carry out its biological function[227, 228]. Within the nucleus, heterochromatin is physically separated from euchromatin. This has been established firmly in several studies that have demonstrated that open, gene-rich genomic loci are found in distinct subnuclear regions to condensed, closed and gene-poor chromosomal domains[229-231] (Figure 1.5). High-throughput chromosome conformation capture technique (Hi-C) has allowed

the identification of chromatin-chromatin interactions genome-wide[232], and revealed that higher eukaryotic genomes are organised into areas of active chromatin (A compartments) and inactive chromatin (B compartments)[232] (Figure 1.5). These compartments consist of a number of super-TADs, TADs and sub-TADs[78, 233-236].

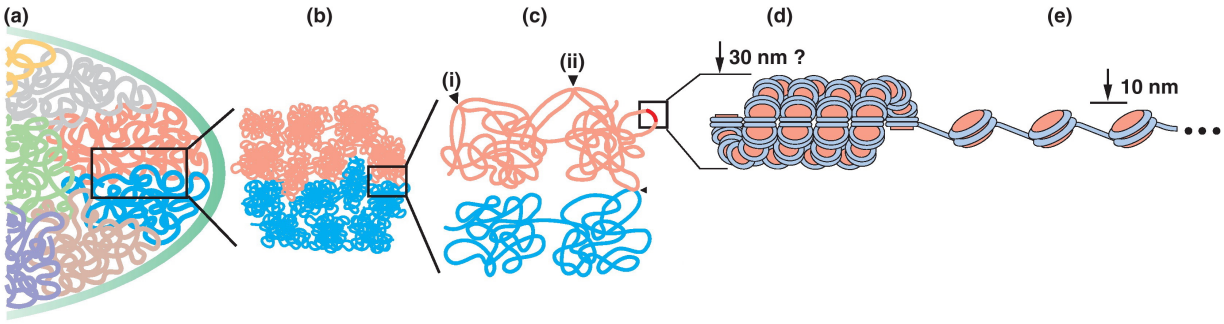


Figure 1.5: Levels of chromatin organisation in the mammalian nucleus

(a) Chromosomes are organized into chromosome territories (CT). (b) CTs are compartmentalised into A & B (Red and Blue globules). (c) Topologically-associated domains (TADs) interact (i) within sub-TADs (frequent), (ii) between TADS of the same CT (rare), or (iii) between close CTs (very rare). (d) Chromatin as a 30 nm fibre, then (e) as a 10 nm series of nucleosomes (beads on a string conformation). Figure adapted from Hubner *et al.* [237].

TADs are largely invariable across different tissues and cell lines, and highly conserved in evolution[238], but recent evidence indicate TAD boundaries can display variable permissibility from 'weak' TAD boundaries that permit more inter-TAD interactions, to the more strict 'strong' boundaries that demarcate adjacent TADs[239]. Exposure to heat-shock in fruit flies resulted in changes in TAD boundaries and merging of neighbouring TADs[240]. Another example comes from the Hox cluster in mammals where TAD/sub-TAD boundaries are not rigid and their fluidity is associated with changes in gene expression during motor neuron differentiation[241]. TAD boundary strength is apparently positively linked to the binding of CTCF[241] (Figure 1.6a).

TADs were first identified in a Hi-C analysis in mouse ES cells which identified 2200 topological domains interacting locally, and occupying over 90% of the genome sequences[235]. These domains displayed enrichment for housekeeping genes, short-interspersed repeat elements (SINEs), and punctuated by boundary elements with insulator characteristics such as the heterochromatic H3K9me3 histone modification and CTCF-binding (Figure 1.6c). The hierarchical architecture of chromatin domains was also found to be conserved between human and mouse, and invariable across

tissues. When compared with previously described A and B compartments described, replication time zones, lamina-associated domains (LADs) and large organized chromatin K9 modification (LOCK) domains, these domains appeared related to, but distinctly different from each of these domain-like structures[235]. Using ChIP-seq in a follow-up investigation, a significant concurrence with *cis*-regulatory enhancer-promoter units was observed in this domain in 19 embryonic and adult mouse cell lines[242]. Genes within a single TAD are under transcriptional coregulation, and whilst TADs generally do not alter during development, intra-TAD interactions vary and rearrange during ESC differentiation, connecting chromatin and transcriptional regulation[78].

TADs partition the chromosomes in such a way that adjacent loci have low interaction frequency if they happen to be located on different TADs. This effect constrains the effect a regulatory element such as an enhancer could exert on the DNA. The exact number and size of TADs remains an open question, with number varying from 2200 originally identified TADs in mice to between 4000-9000 TADs in humans, and from 880 kb in size to between 40-3000 kb respectively[219, 238]. TADs consist of smaller sub-TADs (100s of kb long) and 'contact domains' (10-100 kb long)[219, 234]. CTCF clustering and transcriptional coregulation are correlated with TAD scales during differentiation[243]. Disruption of TADs have been linked to a number of clinical phenotypes such as limb malformations and cancer[244, 245]. Mitotic chromosome with deleted TADs display a more fluid genomic structure and various changes during cell cycle[244]. Higher TAD boundary insulation correlates with increased CTCF levels and varies across tissues. Super-enhancers are observed to be preferentially insulated by strong TAD boundaries, and are commonly co-duplicated in cancer patients[246].

Deletion of the cohesin-loading factor Nipbl in mouse liver leads to significant alteration of chromosomal folding. TADs vanish globally, even when transcriptional changes are not detected[247]. Structural variants can also reshape TADs, resulting in large-scale rewiring of regulatory interactions and gene mis-expression[248]. Interestingly, compartmentalisation is maintained and even reinforced. On the contrary, TAD removal reveals a finer compartment structure that accurately preserve the epigenetic landscape. The 3D organization of the genome is apparently the product of two separate mechanisms: cohesin-independent compartmentalisation of the genome, dictated by chromatin configuration; and cohesin-dependent TAD formation, linking distal enhancers to their target promoters by loop extrusion[247].

As mentioned previously, open and closed chromatin are located in distinct, spatially separated nuclear compartments, A and B compartments. These are further

divided on basis of the pattern of their chromatin-chromatin interactions into A1-2 and B1-4 respectively[232]. Although both sub-compartments of A include highly transcribed genes and active chromatin histone marks (H3K427ac, H3K79me2 and H3K36me3, A1 contains shorter genes with higher GC-content and finishes its replication much earlier than sub-compartment A2[249]. Sub-compartments of the B compartment are found in the periphery of the nucleus near the nuclear lamina, or in the nucleoli, where each exhibits a different feature: facultative heterochromatin (B1), pericentromeric heterochromatin (B2), exclusive association with the nuclear lamina (B3), or only present on chromosome 19 (B4)[219, 232]. Inactive heterochromatic DNA in the B compartments associates with nuclear lamina either directly or indirectly though lamin-associated proteins[250]. Nucleolus-associated domains carry loci mainly transcribed by RNAP I and RNAP III, but they also have a number of RNAP II-transcribed gene from the olfactory receptor family[251]. These genes were intriguingly silenced in the cell-lines where their presence in nucleolar chromatin was observed. This could be a mechanism for silencing RNAP II-dependent genes via chromatin compartmentalisation.

1.1.4 Transcriptional regulation via *cis*- and *trans*-acting variation

Although several layers of gene regulation were explored above, genetic causes of regulatory alterations to gene expression can be broadly categorised into two main types: *cis*- and *trans*-acting variation. *Cis*-acting variation is defined as changes that occur within of physically close to the gene they affect, and include various elements, some of which have already been discussed above such as promoters, enhancers, microRNA binding sites in 3'-UTRs, and splicing variants. On the other hand, *trans*-acting variation instigate their effects on gene regulation via diffusible elements in the nuclear environment that affect physically distant/unlinked genes through the occupancy of their binding sites, such as transcription factors, Mediator protein complex, RNA binding proteins[252].

The binding affinity of a TF provides a useful measure to study the transcriptional activation and the specificity of the spatiotemporal pattern of its binding to gene regulatory regions[253-255]. TF binding specificity and intensity is, at least partially, by *cis*- and *trans*-acting variation, but our understanding of how it occurs is still lacking[256]. Unravelling the interplay between TF occupancy and the other components of the gene regulation machinery is key to understanding phenotypic diversity. There are mainly two approaches to dissect *cis*- and *trans*-acting influence: expression quantitative trait loci (eQTL) analysis and the use of F1 crosses from

genetically inbred organisms[257]. The first approach, eQTL, is based on the correlation of a measured molecular trait, in this case gene expression or TF binding affinity, with sequence variants. The second, F1 hybrid crosses, investigates regulatory mechanisms by analysing the allele-specific pattern of divergence occurring in F1 hybrids compared to the parental strains. The placement of two alleles from different genetic backgrounds into a shared nuclear environment and comparing their relative allelic binding, the extent to which they are influenced by *cis* and *trans*-effects can be measured and evaluated[258]. More on these methods will follow in section 1.4.4.

A 2015 study compared the splicing differences between cultured fibroblasts derived from the inbred mouse strains C57BL/6J (*Mus musculus domestics*) and SPRET/EiJ (*Mus spretus*) to investigate the extent of *cis*- and *trans*-regulatory contributions genome-wide[259]. They investigated the allele-specific splicing patterns in the F1 hybrid of the two mice species, and found that 417/5802 alternative splicing events (~7%) were differentially regulated between the two F0 parental species. 381 (6.6%) of these events showed allele-specific patterns in the F1 hybrids. The parental splicing divergence was found to be the result of *cis*-acting regulatory variation (255 significant *cis*-influenced divergence compared to 62 significant *trans*-influenced divergence). Further analysis of liver tissues in mice F1 hybrid strains showed the same pattern of gene regulation in gene expression and TF binding predominantly via *cis*-acting variation[257, 260]. Observations in *Drosophila melanogaster* were; however, different. A study in 2014 reported that *trans*-acting variants played an equally important role in splice-site divergence[261].

Using F1 hybrids to study *cis* and *trans* regulatory effect has proved invaluable to determining the relative importance of these changes in different gene regulatory layers. Combining this approach with QTL studies in humans, where data from all main levels of gene regulation and expression, could provide a fantastic opportunity to study the genetic basis of phenotypic variability in health and disease[252]. These approaches enable us to understand quantitatively the mechanism by which the different levels of regulatory variation contribute to tissue-specific transcriptional regulation[257].

1.2 The CCCTC-binding Factor, CTCF

CTCF is an architectural protein that is generally considered a master regulator of genome state and function. Loop domains, TADs and chromosome compartments are all enriched for CTCF binding sites with a highly sequence-specific, information-rich motif[170, 262, 263]. The molecular structure of the protein consists of 11 zinc fingers,

in complex with zinc ions bound to cysteine and histidine residues, forming the central DNA-binding domain, surrounded by loose C- and N-terminal domains[264] (Figure 1.7a). CTCF is suggested to act through recognition of diverse DNA sequences by combinatorial usage of its 11 zinc-fingers, a sort of "CTCF code"[265]. CTCF is an essential cellular protein, and its 11 zinc-finger DNA-binding domain is highly conserved in higher eukaryotes, with a 99% and 98.7% amino acid identity between humans versus chicken and Xenopus frogs respectively[266-268]. CTCF activity regulates gene expression in various way: transcriptional activation/repression, enhancer blocking and setting up boundary elements/insulators, tethering promoters and enhancer and promoting long-range interactions, as well as blocking the spread of active chromatin and demarcating active from silent genomic regions[170, 269, 270]. CTCF bridges the gap between spatial organization of the genome and its function and the underlying gene regulatory processes[271].

The CTCF gene is also conserved in most bilaterian metazoan, widely expressed in both adult differentiated tissues and during embryonic development[266, 268, 272-276]. CTCF is the only insulator protein to be identified in vertebrates, with an enhancer blocking activity *in vivo*[277-286]. In addition to loop domains, TADS and chromosome compartments, CTCF is found in interaction with nuclear lamina and chromatin boundaries, signifying the role CTCF plays in boundary formation. CTCF also interacts with RNA, and many CTCF sites are not engaged in chromatin folding or loop formation[287, 288]. CTCF genomic occupancy is sensitive to the methylation status of DNA, providing a measure of control on the regulation of epigenetically imprinted gene expression[289-293]. Mis-regulated DNA methylation in cancers with metabolic deregulation disrupts CTCF binding[294]. Chromatin conformation assays have demonstrated that the presence of CTCF binding sites correlates with long-range interaction[39, 235]. CTCF binding is additionally reported at the transitions between distinct chromatin states, marked by histone modifications[295]. This supports the hypothesis that at least a subset of CTCF sites are capable of forming boundaries, besides blocking the spreading of regulatory effects. There are; however, many more CTCF sites bound *in vivo* than chromatin boundaries[269], indicating the wide-range of functions CTCF is naturally involved with.

The following section will focus on the role of CTCF in gene regulation and chromatin structure. It explores the basis of CTCF binding and its binding site, and the functional consequences on gene regulation, and how it associates with the cohesin protein complex to achieve its wide-ranging activities in genome organisation and function.

1.2.1 CTCF binding: features and consequences

CTCF binds with a high-affinity to a nonpalindromic canonical motif in its binding site with a sequence consensus referred to as M1[269, 296-299] (Figure 1.7a). Studies have shown that the central zinc fingers, 4-7/8, are needed for this interaction[300]. This 20 bp core motif is common to nearly all known CTCF binding sites as identified by various immunoprecipitation methods, and the involvement of nonspecific zinc fingers, other than the ones mentioned previously, with the surrounding DNA sequence helps stabilize the binding[297]. A second 10-bp motif, termed M2, is found upstream of the canonical M1 separated by a DNA spacer[269, 297, 301], where it interacts with the 9-11 zinc fingers[302]. Findings suggest that the M2 motif is in conjunction with the M1 in 15%-25% of all CTCF binding sites, whereby CTCF binds with high affinity depending on the spacer between the motifs[303].

The presence of a CpG in the canonical motif's consensus sequence lends support to the idea that methylation of cytosine residues at carbon 5 of the base to form 5-methylcytosine (5mC) in CGI-harbouring CTCF binding sites may underlie CTCF selectivity in different cellular contexts[304]. Studies support a model where DNA methylation is a common regulatory measure to control CTCF occupancy at many loci, such as CDKN2A, B-cell CLL/lymphoma 6 (BCL6) and brain-derived neurotrophic factor (BDNF)[305-307]. Comparison of DNA methylation patterns in 19 human cell lines with mapped CTCF occupancy showed that 41% of tissue-specific CTCF binding sites are linked to differential DNA methylation[293]. On the other hand, 67% of those sites that were linked with variable DNA methylation, the presence of 5mC correlated with a corresponding downregulation of cell-type-specific CTCF binding. CTCF also forms a complex with poly(ADP-ribose) polymerase 1 (PARP1) and DNA (cytosine-5)-methyltransferase 1 (DNMT1), activating PARP1, which in turn inactivates DNMT1 by poly(ADP-ribosyl)ation, maintaining methyl-free CGIs in the genome[308, 309]. Furthermore, other studies in mammals have observed that CTCF can also cooperate with RNAs to stabilize its interactions with other protein complexes, such as the DEAD-box RNA helicase and p68 and their associated ncRNA[310]. These, along with more recent findings that demonstrate that CTCF binds to the *Jpx* RNA, indicate that ncRNA are involved in the stabilising of interactions mediated by CTCF and its protein partners[311]. Saldana-Meyer *et al.* reported at least 17,000 genomic RNAs that interact with CTCF[287].

One of the most interesting findings of recent years is that a pair of CTCF binding sites will only engage to fold chromatin, forming long-range loop interactions if they are in a convergent, linear orientations, producing asymmetrical insulator pattern[219, 312]. The inversion of a single CTCF site is sufficient to rewire the

orientation of the looping and disrupt the packaging of the underlying chromosome segmentation pattern into an insulated TAD, proving that the proper arrangement of binding sites is crucial for the correct functioning of CTCF[244, 279, 282, 285]. In addition, the deletion of a TAD boundary in the vicinity of the Xist locus on chromosome X results in ectopic loop interactions general mis-regulation of gene expression[78]. Analysis of the Six homeodomain locus in zebrafish unveiled CTCF binding sites in oriented convergently with TADs at TAD boundaries, and attempted deletion of any of these boundaries results in erratic interdomain enhancer-promoter interactions[313].

Other regulatory factors may also contribute to augmenting or modulating CTCF function[314]. For example, Smad proteins interact with CTCF at the Igf2/H19 imprinted control region[315]. Similarly, at the Igf2/H19 locus, p68 helps, along with the long noncoding RNA SRA, to stabilize cohesin binding and create an effective insulator. DNA-bound CTCF/cohesin complexes recruit the core promoter factor TFIII to helps stabilize CTCF binding at specific promoter-proximal regions at many loci in ESC[316, 317]. CTCF also associates with PARP1 to establish inter-chromosomal contacts during the circadian cycle[318].

Homozygous knockout of CTCF is embryonic-lethal[319-321], and partial deletion of CTCF leads to an altered gene expression pattern, yet with more limited phenotypic impact, increasing radiation sensitivity, defective DNA-repair mechanism, and cell cycle arrest[80, 322]. Full removal of CTCF results in total loss of nearly all loop interactions in a highly dose-dependent manner[247, 323, 324]. Conditional *Ctcf* knockouts in a tissue-specific context, such as in oocytes, lymphocytes, neurons, and cardiomyocytes, lead to organ failures[276, 325-327]. Acute depletion of CTCF *in vitro* by both RNAi and transient auxin-mediated in mouse ESC yields full removal of CTCF from the nucleus, disruption of loop structures and TADs, yet high-order chromosome compartmentalization is maintained[269, 323, 328]. Although Ctcf hemizygous mice undergo normal development, they exhibit an increased predisposition to tumours[329]. Even though halving of CTCF protein concentration is physiologically tolerated, the process reduces the overall fitness of the organism. CTCF has also been shown to be a haploinsufficient tumour suppressor gene in human cancers[329-331]. A recent study observed that Ctcf hemizygous cells show modest but consistent changes in almost 1000 CTCF binding sites that are of lower affinity and weaker evolutionary conservation across the murine lineage. This coincided with dysregulation of several hundred genes' expression, which are ontologically enriched in cancer-related pathways. Chromatin configuration is, however, unaffected apart from disruption to some loop domains[332]. Mutations of CTCF motifs lead to oncogene dysregulation in some cancers[294], and defective limb development in humans and

mice[244]. Unlike germline variants, somatic missense and nonsense mutations of CTCF are abundant in human tumours [333, 334]. Hyper-methylation of the GC-rich CTCF binding motif was observed to decrease CTCF occupancy in glioma, and constitutive CTCF–CTCF binding site interactions are reportedly deleted in T-cell acute lymphoblastic leukaemia, resulting in oncogenic upregulation[245, 323].

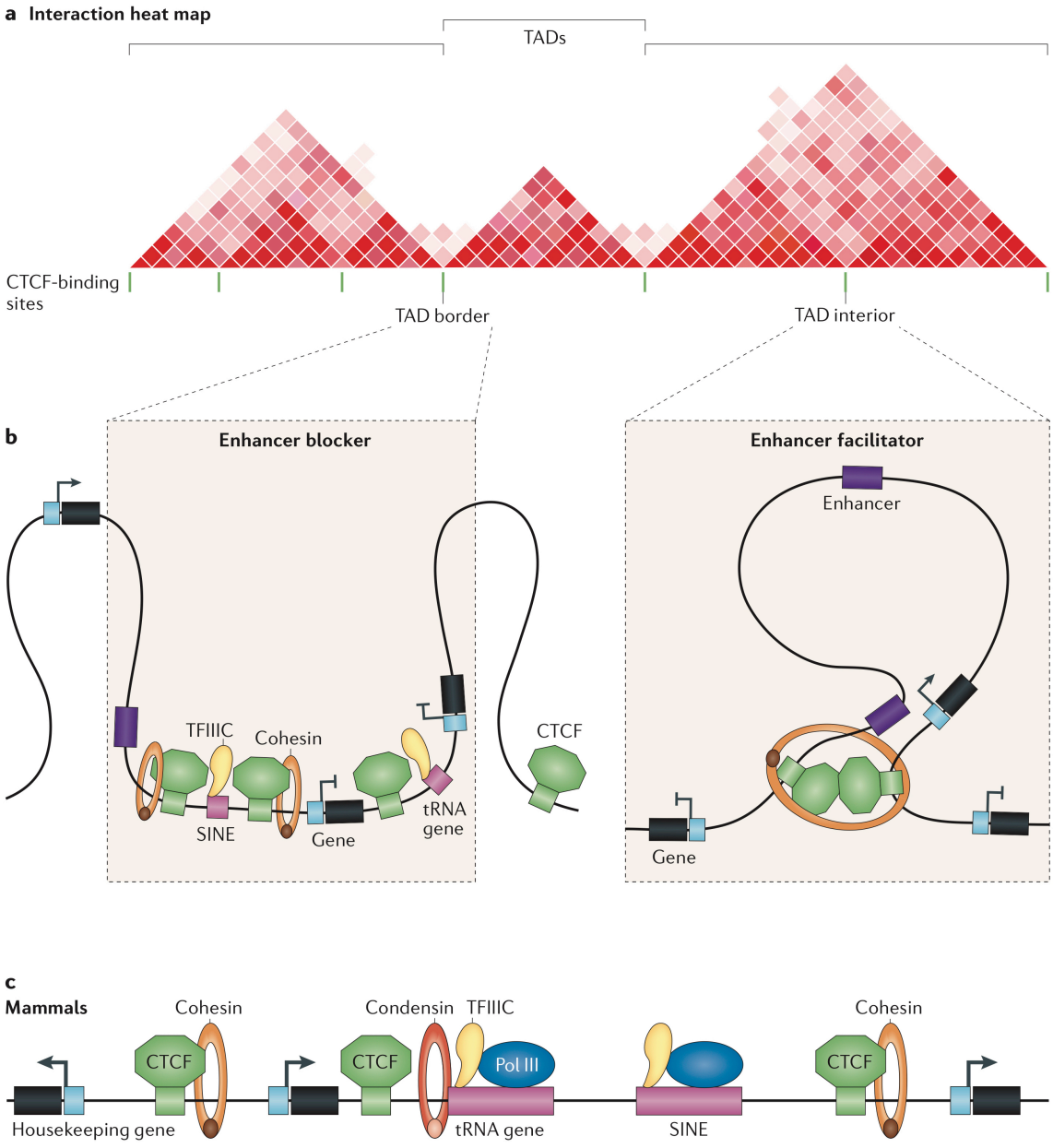


Figure 1.6: CTCF regulates 3D chromatin architecture

(a) An interaction heat map of a ~2.5-Mb chromosome segment. TADs, their boundaries and interiors are indicated. (b) multiple CTCF binding sites and TFIIIC

at TAD borders contribute to its formation. CTCF may act an as an enhancer blocker (left). On the other hand, CTCF bound inside TADs may act as an enhancer facilitator through looping the DNA with the help of cohesin. Blue boxes denote gene promoters, and black boxes denote genes. (c) The TAD borders in mammals are enriched for housekeeping and tRNA genes, SINEs and CTCF-binding sites. Figure adapted from Ong and Corces [80].

Intriguingly, Satou *et al.* recently found that CTCF binds to a motif in the Human T lymphotropic virus type 1 (HTLV-1; the human T-cell leukemia virus), when HTLV-1 is inserted into the host cell genome[335]. It is hypothesised that CTCF binding to the provirus can promote abnormal chromatin looping by dimerizing with CTCF in the surrounding host genome. The presence of a single CTCF-dependent chromatin loop *in vitro* T cell line has since been demonstrated[335, 336].

1.2.2 CTCF roles and functions

CTCF is a versatile nuclear factor, involved in various roles such as transcriptional activations/repression, insulation, regulation of genetic imprinting, developmental programme modulation, structural domains organization and guarding genomic fidelity[337] (Figure 1.6). The mechanisms underlying the diverse functions of CTCF in genome biology derive from its function in mediating long-range interactions between two or more DNA sequences. 4C analyses of the mouse imprinted maternal H19–insulin-like growth factor 2 (Igf2) locus demonstrated that the H19 imprinting control region (ICR) is involved in extensive inter-chromosomal and intra-chromosomal interactions across the genome that require the CTCF binding within the ICR[338]. CTCF binding at several DNase I hypersensitive sites is central to preserving the unique chromatin architecture at the murine haemoglobin subunit beta (Hbb) locus[339]. CTCF-mediated interactions modulate facets of genome function in a context-dependent manner. The functional consequences of these interactions rely on the sequences flanking CTCF- binding sites and the presence of other specific architectural proteins[80].

Furthermore, CTCF promotes transcriptional activation of some genes, such as the case of CTCF binding to the amyloid precursor protein (APP) promoter[340] (Figure 1.6b). The structural domain of 107 amino acids in the N-terminal tail of CTCF regulates transcriptional activation and chromatin de-condensation, and upregulates its expression as it approaches the promoter location[341]. Conversely, CTCF may also play a role in transcriptional inhibition, by combining promoter and upstream silencer together. CTCF was originally identified as a transcriptional repressor of chicken *c-Myc* gene[266, 342]. CTCF binding along with thyroid hormone

receptor to the isogenous locus forms a repressor complex that lead to c-Myc reduced expression[343]. CTCF-mediated transcription repression could be achieved via recruitment of histone deacetylase and deacetylation of CTCF via binding of SIN3 transcription regulator family member A[344]. Recent genome-wide studies indicate that CTCF can additionally act as an enhancer blocker at particular loci. 15,000 CTCF binding sites were identified in human genome-wide search for conserved regulatory motif. These sites appeared to demarcate adjacent genes which show notably conversed correlation in gene expression compared with genes that are in a similar architecture, but that are not separated by CTCF-binding sites[299]. CTCF also works as an insulator bounding factor inhibiting interactions between promoter, enhancer, and silencer, provided that the CTCF binding site resides in-between regulatory elements that fail to properly function[345]. A study identified a 42-bp insulator sequence that could block the promoter activity of β-globulin, and equally works as a binding site of CTCF in humans[346].

Despite being originally thought off as an insulator and blocker of gene activity(Figure 1.6b), recent studies have identified CTCF as an important factor in tethering distant enhancers to their promoters. 79% of long-range interactions between promoters and their regulatory sequences were shown not to be blocked by the presence of one or more intervening CTCF-bound sites[39]. Strikingly, a subset of these long-range interactions are significantly enriched for CTCF and/ or histone modifications that are marked for active enhancers such as H3K27ac, H3K4me1 and H3K4me2. These results propose an alternative role for CTCF in genome biology may be to facilitate the communication between regulatory elements and promoters. Further support for this hypothesis came by finding a significant overlap between tissue-specific CTCF occupancy and enhancer elements, in addition to similar studies at several other loci[242]. For example, activation of major histocompatibility complex class II (MHC-II) gene expression by treatment with interferon-γ (IFNγ) requires CTCF-mediated looping of the XL9 enhancer element and its core promoters, MHC class II transactivator (CIITA) and specific transcription factors[347]. Thus, CTCF-mediated topological organization precedes transcriptional activation[348].

CTCF is also involved in regulating transcriptional pausing and modulating alternative mRNA splicing. For example, the first intron and upstream regulatory sequence in the mouse myeloblastosis oncogene (Myb) locus are bound by CTCF. CTCF-mediated looping between the first intron, promoter and upstream enhancer elements, along with its associated erythroid transcription and elongation factors is necessary for RNAP II to mediate transcriptional elongation and the upregulation of the Myb gene during erythroid differentiation[325]. The genome-wide distribution of CTCF binding sites at promoter-proximal regions and in 5'UTRs clearly correlates

with high pausing indexes suggesting that the effect of CTCF on Pol II elongation may be more common than previously thought[349]. Recent studies have shown that disruption of mRNA elongation by RNAP II by CTCF may cause the inclusion/exclusion of particular exons in the mature mRNA[350, 351]. CTCF binding to exon 5 of CD45 gene promotes its alternative splicing in the mRNA, whereas blocking CTCF binding results in removal of this exon from the final edited mRNA[351].

An additional role of CTCF is in chromosome X inactivation. During development in mammals, one copy of the X chromosome's pair in females undergoes a process of inactivation as a measure of gene-dosage control. The process relies on the expression of the inactive x-specific transcript (Xist) and is inhibits by the antisense gene Tsix[352]. The imprinting centre of the X chromosome harbours a battery of CTCF binding sites with methylation-sensitive enhancer blocking activity. CTCF in association with the inhibitory Tsix regulates the epigenetic switch of X chromosome inactivation by stimulating Tsix transcription or blocking Xist from interacting with its enhancer. Expression of Tsix prevents Xist mRNA accumulation[352].

1.2.3 CTCF and Cohesin

CTCF and the ring-shaped cohesin complex have been repeatedly shown to colocalize in the genome[353] and bind at the anchors of chromatin loops[219, 339], and the TAD boundaries[78, 232, 235], demonstrably indicating their critical involvement in regulating genome folding. Targeted deletion of these sites disrupted loop formation and contact domain structures[279, 282, 285].

Cohesin is an architectural protein complex in the shape of a large ring molecule. Similar to the highly related condensin and Smc5/6 complexes, the cohesin complex core is made up of heterodimers of structural maintenance of chromosomes proteins (SMC), a highly conserved family of ATPases. The V-shaped SMC1-SMC3 heterodimer is formed when they join their coiled-coil and hinge domains. This is then complemented by RAD21 (also known as SCC1), and the addition of SA1/SA2 (also known as STAG1 and STAG2) subunits to complete a 'ring' structure large enough to topologically 'embrace' two chromatin fibres (Figure 1.7b). Other proteins that associate with the complex include the SCC2/4 (also known as Nipbl/Mau-2 in mammals) adherin complex that help cohesin loading onto chromatin, and WapI, which is required for eventual cohesin removal[354, 355]. Cohesin is a major component of chromatin in cycling, non-cycling, and post-mitotic cells in higher eukaryotes.

Cohesin was originally identified as a complex that provides cohesion to the two sister chromosomes during DNA replication in S phase until cell division[87], allowing post-replicative DNA repair and faithful chromosome segregation and the fidelity of genomic information passed on from mother to daughter cells (or from a generation of multicellular organisms to the next) in both mitotic and meiotic cell division modes[224, 356]. In line with its role in post-replicative DNA repair and chromosome segregation, cohesin loading is increased at sites of DNA damage[357, 358] and at centromeres[354]. pointing to a role for cohesin outside of the division phase of the cell cycle. Accumulating evidence indicates that cohesin does play a role in mediating chromatin structural conformation and gene expression in interphase. Although cohesin partners with CTCF to play it its role in transcriptional regulation, it can also function independently from CTCF to achieve gene regulation in tissue-specific context, for example via loading onto the promoter and enhancer elements in oestrogen-regulated gene expression[359, 360]. Whereas strong cohesin sites overlap with CTCF binding, 'weaker' cohesin sites map onto active promoters and enhancers, where it interacts with Nipbl, the Mediator complex, and cell-type-specific transcription factors to instigate its regulatory potential[224, 359, 361, 362]. In addition to CTCF and cohesin well-established roles in chromatin and transcriptional regulation, further investigations have associated them with various other roles such as transcription factor binding[361], transcriptional elongation[363], alternative splicing[350, 351] and interactions with noncoding RNAs[364].

CTCF and cohesin binding is associated with ~90% of DNA loops, 92% of which involve CTCF anchor in convergent orientation that face the loop interior[219] (Figure 1.7c). This recently identified feature of CTCF loop formation facilitates the prediction of loop formation *in silico*[285, 365]. When these CTCF anchor are inverted or deleted (e.g. by CRISPR/Cas9 genomic editing), the expression of nearby genes changes as predicted[327]. More recently extreme deep sequencing Hi-C studies in mESCs found ~10,000 previously unidentified shorter loops with a median size of 185 kb in a complex nested structure[219]. These short, <200 kb CTCF-anchored loops (termed chromatin contact domains or super-enhancer domains) were enriched for CTCF and cohesin binding sites and tissue-specific genes and enhancers[327, 366]. They are; however, in the minority and do not explain the nuclear topological domains seen in high resolution Hi-C maps[219, 324, 367].

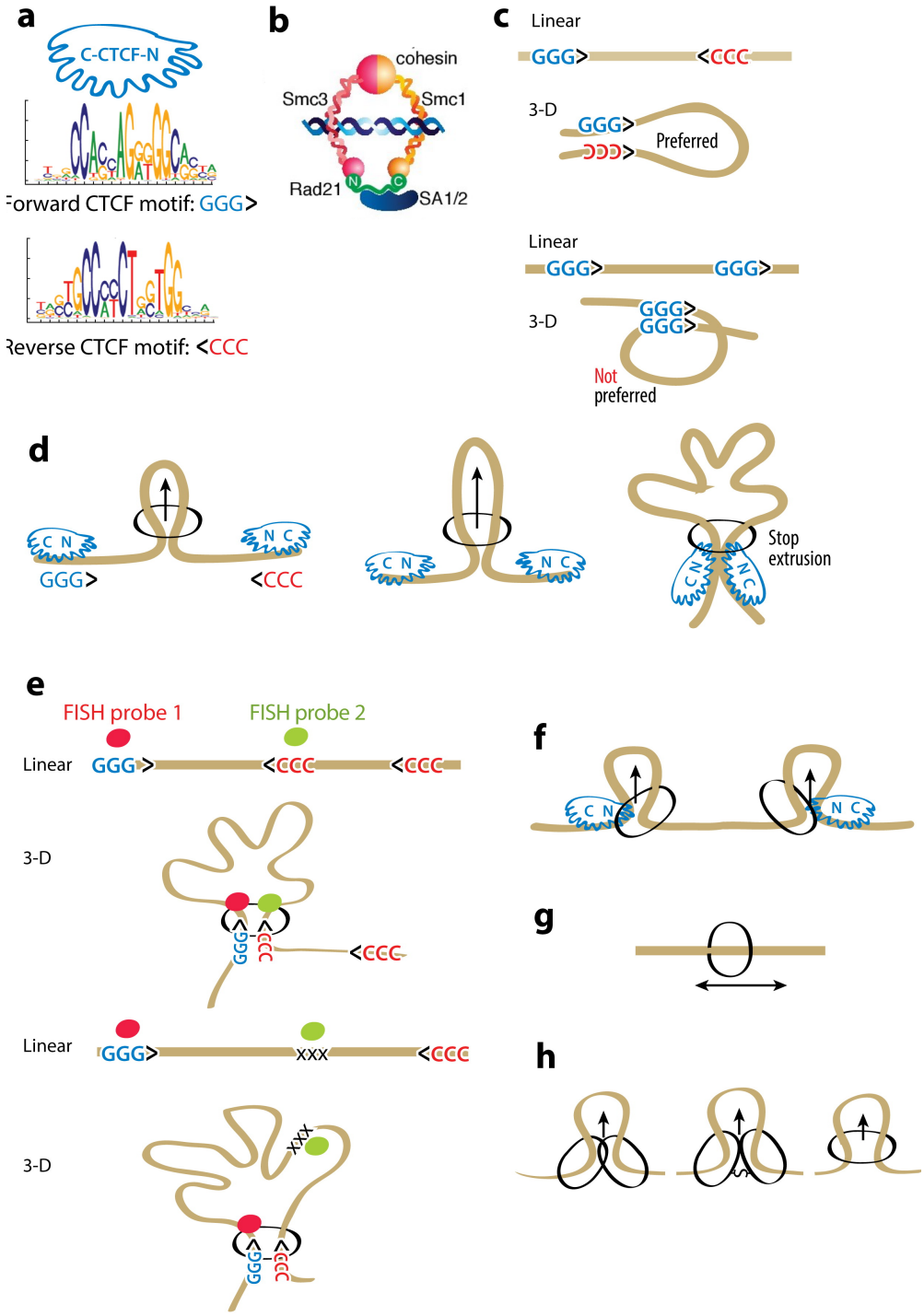


Figure 1.7: CTCF and cohesin are essential for the extrusion model of genome folding

(a) CTCF and its motif. "GGG>" is the forward motif (blue), and "<CCC" is the reverse motif (red). The 11 zinc fingers of CTCF bind the CTCF canonical motif, and the C-terminal fingers project toward the 5' end of the motif. (b) The cohesin complex:

the core subunits SMC1, SMC3, Rad21, and SA1 or -2. The ring-like structure accommodates 2 chromatin fibres. (c) Loops form primarily between convergent CTCF binding sites. (d) Extrusion forms chromatin loops and stops when two convergent CTCF-bound sites are encountered. (e) If a TAD border is lost, extrusion continues to form a larger TADs. The original TAD boundaries are marked with red and green fluorescence *in situ* hybridization (FISH) probes. (f) Simultaneous extrusions form two loops. (g) A single cohesin ring "embracing" a chromatin fibre slides smoothly along the genome without extruding a loop. (h) Two cohesin rings could embrace the chromatin fibres to extrude a loop by being linked either topologically (left) or by complex formation (centre), or a single cohesin ring sliding from the top of a preformed chromatin loop (right). Figure adapted from Merkenschlager and Nora & Hansen *et al.* [355, 368].

The extrusion model provides the mechanistic framework to explain how CTCF and cohesin could regulate chromatin topology and the underlying gene expression (Figure 1.7d). As the model proposes, when cohesin loads onto a genomic region in the vicinity of properly oriented CTCF anchor, an extrusion subunit is arrested, while its partners proceed along the chromatin. The CTCF-anchored element rapidly and transiently interacts with the entire genomic interval[369]. Recent work has shown that such a mechanism can facilitate V(D)J recombination at the immunoglobulin locus in developing lymphocytes[370]. *In vitro* studies illustrated how RNA polymerases can push cohesin rings along DNA[371]. The asymmetric nature of this model could explain the ability of promoters to interact with enhancers spread across 100s kb. Hansen *et al.* demonstrated how CTCF and cohesin can form a rapidly exchanging 'dynamic complex' rather than a typical stable one, suggesting that chromatin loops are dynamic and in continuous equilibrium between breaking and forming throughout the cell cycle[368] (Figure 1.7f-h).

Although depletion of CTCF decreased cohesin occupancy at their binding sites, cohesin loading onto chromatin is unaffected. Conversely, the depletion of cohesin complex proteins does not significantly impact CTCF genomic occupancy patterns. Taken together, these observations suggest that CTCF alters cohesin genomic distribution, but not its association with chromatin in general, and acts upstream of cohesin to ensure its proper positioning in relation to its target sites[293, 372]. Experimental knockouts of the SA1 subunit of the cohesin complex, which normally interacts with the C-terminus of CTCF, cause the redistribution of the SMC1/3 subunits and reduce cohesin's association with CTCF binding sites[373, 374].

1.3 The evolutionary genomics of transcriptional regulation

Since Charles Darwin famously put forward his theory of evolution by means of natural selection[375], he forever changed the science of biology in all its branching complexity, and genomics is no exception. Comparative genomics attempt to explain the biological function using an evolutionary perspective, and how phenotypic differences observed in the natural world between species can be understood in the context of the genome. Part of heritable phenotypic variation is due to differences in transcriptional regulation, which determines the extent of gene expression in the different cells and tissues[376, 377].

Gene expression levels are quantitative traits subject to evolutionary processes. Protein-coding genes have been subject to strong selective pressures as revealed by interspecies comparisons of mammalian genomes that confirmed the identity of almost all coding sequences[378]. The genetic basis for gene expression variation must then lie in the non-coding regulatory genome. Experimental evidence in molecular evolutionary studies have steadily enhanced our understanding of the underlying processes that may govern the evolution of gene regulation, especially in the domain of non-coding genome and role of epigenetics. Results from those studies have helped explain phenotypes from the pigmentation in fish and malaria resistance in wild primates[379, 380].

A major mechanism for the evolution of transcription regulation comes from the domain of transposable elements (TEs). It was the work of Barbara McClintock on maize[381] that demonstrated how TEs can control gene expression, and paved the way to recognising their evolutionary role in rewiring the gene regulatory networks[382, 383]. Ever since, TEs have repeatedly been reported to harbour functional TF and DNase I hypersensitivity sites[384-387]. Data from the ENCODE project revealed that 44% of open chromatin regions in the human genome are in TEs, as well as 63% of primate-specific gene regulatory elements, where a particular class of TEs, endogenous retrovirus-like elements (ERVs), have expanded into hundreds of thousands of novel regulatory elements, and reorganised the human gene regulatory landscape[388]. The incredible pace through which repetitive and other fast-evolving sequences evolve explains their ability to alter transcriptional regulatory circuits via the creation and disruption of sequence motifs[389]. This is the reason why mammalian genomes are set apart from other higher metazoans like birds and arthropod whose genomes are depleted from repetitive elements and show more signs of evolutionary constraint[390, 391].

For example, a study used a transchromosomic mouse strain, one that carries an almost complete single copy of human chromosome 21 via the female germline, to reveal that in a heterologous (mouse/human) regulatory context, transposon-derived human regulatory regions become transcriptionally active[392]. Hundreds of loci on the human chromosome 21 became associated with changes in DNA methylation at CpG dinucleotides and histone marks specific for transcriptional activation in germline and somatic tissues, resulting in apparent gene expression of the nearby loci. These sites on chromosome 21 were found to be enriched with primate and human lineage-specific transposable elements[392, 393]. A seminal work by the same group using ChIP-seq to profile the genome-wide occupancy pattern of 2 TFs, CCAAT/enhancer-binding protein alpha (CEBPA) and hepatocyte nuclear factor 4 alpha (HNF4A) in the liver of 5 vertebrates[394]. Although each TF was found to bind highly conserved DNA motifs, most binding events observed were lineage- and species-specific, and highly conserved binding events present in all five species were very uncommon. Genes whose expression levels are TF-dependent display evolutionary conservation and found bound by the TF in multiple species with no increased motif constraint. Motif sequences changes generally explained binding site divergence between species[394].

In this section, I will give a broad view of the current state of research into the evolutionary genomic of *cis*-regulatory elements, with a particular focus on CTCF, and the role TEs play in the process of shaping the regulatory landscape of gene expression, particularly in mammals.

1.3.1 Evolution of *cis*-regulatory elements and TF binding

Whilst protein coding mutations have been well-catalogued and characterised, changes to *cis*-regulatory sequences have only recently come to the forefront of gene expression research. Despite this, only a fraction of those interactions in a limited number of organisms, in selected tissues and developmental stage, and under specific conditions, are known. Alterations to *cis*-regulatory elements can change the course and pattern of gene expression and pave the way for the evolution of species-specific traits[395, 396].

Most components of the transcriptional machinery that regulate gene expression are highly conserved in evolution. As outlined earlier in 1.3, TFs and the sequence of their binding motifs are mostly conserved between human and fruit flies[397-399], and comparisons of the sequence of TFs motifs across different lineages yield a high degree of similarity[394]. It is the *cis*-regulatory elements, such as enhancers whose locations and activities in orthologous sequences that are less

conserved. An examination of the activation profiles in 41 pairs of conserved regulatory elements between human and zebrafish revealed that only a third of these regions displayed any form of conserved activity between the two species[400]. A major study by Villar *et al.* discovered that turnover of the regulatory activity is pervasive between even more closely related species (Figure 1.8). For example, only 1% of human liver enhancers exhibit consistent activity across the 20 mammals investigated[401]. Taken together, these studies suggest that orthologous regulatory regions display a varied level of activity across species, even though the TFs they bind and their motifs are very much conserved in structure.

It has been suggested that the evolutionary conservation of genes and their regulation is a product of pleiotropic trade-off[399, 402-406]. Pleiotropy is the ability of a single gene/variant to influence several traits simultaneously. If a pleiotropic region is altered by mutations, newly-created variants may confer different effects on the multiple functions they contribute to. These variants may be advantageous in one facet, but gravely deleterious in others[406]. Intuitively, these regions should be under selective pressure to remain more evolutionarily constrained than other, non-pleiotropic regions, and it follows that genes with pleiotropic functions are more likely to be found in orthologous regions in other species, and with a comparable expression level[403, 404]. Therefore, TF binding sites that are active tissue-wide and observed at several developmental stages are expected to be conserved between species[399]. Nevertheless, the sequences TFs bind to vary in terms of their information content which correlates strongly with how their occupancy evolves. Genetic drift causes low-affinity, information-poor motifs to evolve rapidly[6, 394, 407-410], but sequence change alone is incapable of fully explaining the evolutionary trajectory of TF binding[407, 408, 411]. Larger, information-rich motifs, such as the CTCF motif, are selectively conserved[269].

Increasingly, investigations into the evolution of mammalian transcriptional regulatory elements are documenting the rapid turnover of enhancers and tissue-specific TF binding sites[412-416]. Findings demonstrate that gene expression across similar tissues in different species is more correlated than different tissues within the same species, suggesting that tissue-specific expression pattern are evolutionarily stable[417, 418]. How this is achieved and maintained in the face of the ever-changing regulatory landscape is a central dilemma in evolutionary genomics. The evolutionary status of regulatory elements ranges from highly conserved to lineage-specific. Taken together, these various findings support the notion that the greater functional influence of a regulatory element, the more conserved across different species, tissues and developmental stages[401, 414, 419, 420]. Conversely, lineage-specific elements appear

to partially compensate for proximally lost events, and are often found in regions with pre-existing regulatory activity[394, 421].

An integrated analysis of transcriptional circuit evolution across >25 animal species examined mRNA expression, transcription factor binding and *cis*-regulatory motifs. The results revealed that transcriptional regulatory networks evolve at a constant rate across the various lineages, even more so when only chromatin-accessible regions were considered[422]. Another more recent study compared conserved-activity enhancers to species-specific-activity enhancers using liver enhancers in ten diverse mammalian species. Conserved-activity enhancers exhibited greater regulatory potential and activity in humans than their species-specific counterparts. They appeared active across more cellular contexts and the genes they regulated were expressed in more tissues, providing further support to the pleiotropy argument mentioned earlier[423]. Bertholet *et al.* followed up their 2015 study by analysing promoter and enhancer activity with corresponding gene expression levels in liver samples from 15 species, and reached similar conclusions[424]. They also reported that the evolutionary resilience of transcription is dependent on the number of regulatory elements, with an emphasis on evolutionary conservation. Elements with conserved activity in more species have the most ability to drive stable gene expression. Recently-evolved species-specific enhancers, on the other hand, have a weaker overall regulatory potential[424].

The mechanisms of gene regulation can be influenced by *cis*- and *trans*-acting variation with local and pleiotropic effects, respectively. These changes can instigate a much wider effect resulting in changes to gene regulation evolutionary dynamics[425]. A study used the analysis of RNA-seq to measure liver gene expression divergence between two mouse strains: C57BL/6J (*Mus musculus domesticus*) and CAST/EiJ (*Mus musculus castaneus*) to establish the extent of allele-specific expression in their F1 hybrid offspring. 535 genes were identified which displayed a parent-of-origin-specific patterns of gene expression, but only few of those genes suffered complete allelic-silencing, indicating that genetic imprinting in somatic mouse tissues accounts for a relatively small number of genes[260]. 32% of non-imprinted genes demonstrated divergent expression between the parental F0 strains., of which only 2% were found out to be exclusively influenced by *trans*-acting variants. 43% of the set of non-imprinted genes were attributed to variants acting only in *cis*. The remainder of genes (55%) showed gene expression divergence pattern consistent with a combinatory complex of *cis*- and *trans*-acting variation. The genes whose expression divergence is driven by *trans*-acting variation were additionally observed to have higher sequence constraint than genes whose divergence was caused by variants acting in *cis*. Gene expression changes instigated by variation in *cis* and *trans* were interestingly in

opposite directions, suggesting that compensatory regulation due to purifying selection may work to stabilize gene expression levels[260].

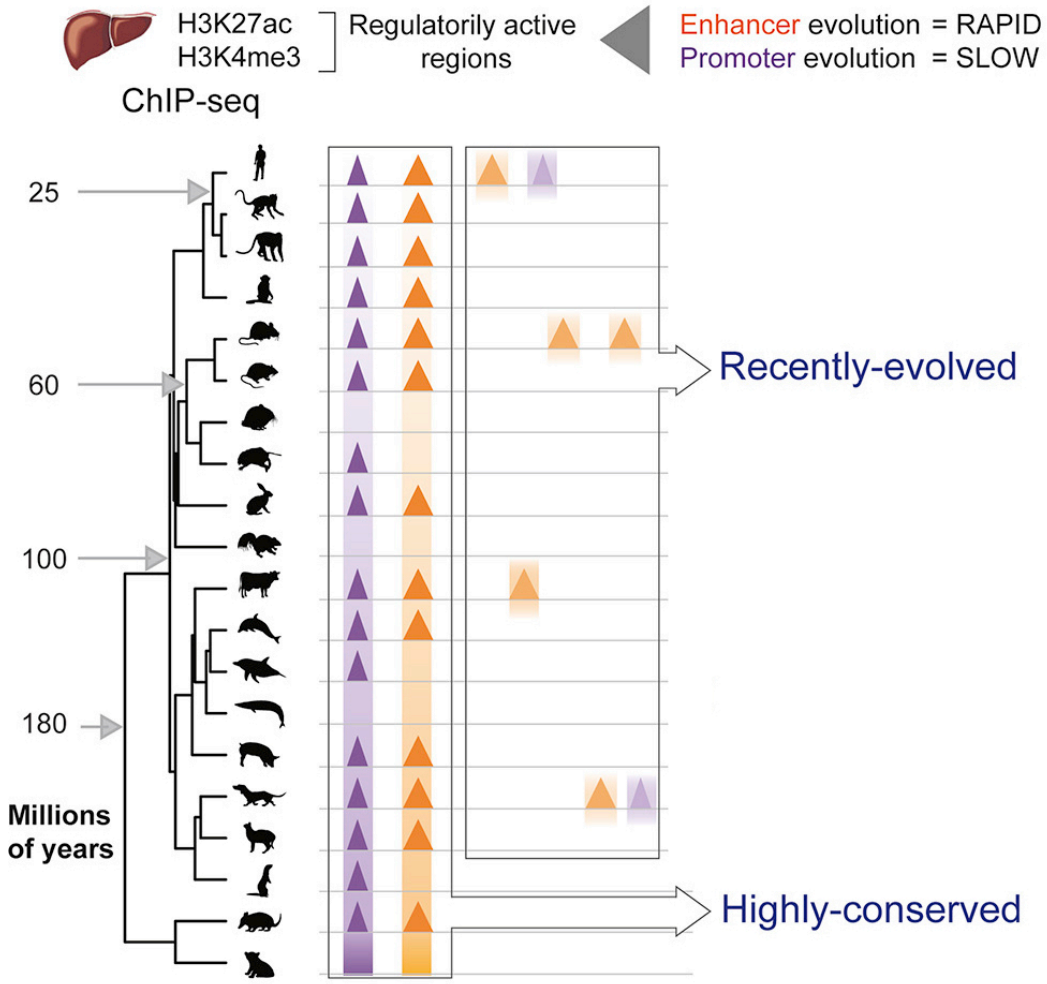


Figure 1.8: Enhancer and promoter evolution in 20 mammalian species

Comparative evolutionary genomic analysis in 20 mammals reveals rapid enhancer (orange triangles) and slow promoter (purple triangles) evolution across the evolutionary tree. Enhancers are only rarely constrained, and recently-evolved enhancers are predominant in their regulatory landscape, exhibiting lineage- specific positive selection. Grey arrows on the left indicate divergence time since last common ancestor in millions of years. Figure adapted from Villar *et al.* [401].

A more recent effort investigated tissue-specific TF (CEBPA, HNF4A, FOXA1) occupancy divergence in the livers of the same two strains and their F1 hybrid to highlight the contributions of *cis* and *trans* variation on the dynamics of gene regulation[257]. They also identified that *cis*-directed mechanisms are

predominant in the birth of new TF binding sites in lineage-specific manner. Furthermore, they detected apparent coordination in the regulatory networks between TF occupancy, chromatin state, and gene expression in the F1 hybrids[257].

1.3.2 Evolution of CTCF binding

The multifaceted nature of CTCF roles have placed a strong purifying evolutionary pressure on its binding sites[298, 299, 384, 409, 426, 427]. Evidence from previous studies indicates that CTCF binding is evolving at a rapid pace despite the selective pressures it is under. Remarkably, the genome contains thousands of CTCF binding sites are found in rodent-specific SINE B2 repeat elements, meaning that they are not conserved with their human counterparts[384]. This provides an excellent example of a case where the early models of TEs as modulators of gene regulatory evolution via expansion of repeat elements driving divergence in eukaryotic genomes[409, 427-432].

The link between CTCF binding site evolution under the influence of SINE TEs first arose in a study of TFs occupancy patterns and their association with TEs[384]. A substantial portion of the B2 SINE TEs carrying the CTCF binding motifs in the mouse are bound by CTCF *in vivo*. This mechanism offers the means for rapidly expanding the long complex CTCT motif into a multitude of novel sites. This mechanism had previously been proposed for the repressor REST/NRSF, which also has a similarly large and complex binding motif[433]. A landmark study comparing CTCF motif occupancy in six mammals established that SINE repeats, which are incidentally still active in multiple mammalians lineages, carry the canonical CTCF motif[269].Hundreds to thousands of such sites were identified in dog, opossum, rat and mouse. The sequence surrounding the CTCF sites that are the oldest and most conserved are enriched for hundreds of fossilized SINE TEs in several mammalian species, separated by 180 million years of evolution. The various findings support an ancient mechanism of genome evolution, based repeat-driven expansion of CTCF binding sites from a set of ancient sites to their current genomic distribution. Primate genomes, remarkably, seem to have escaped this mode of regulatory rewiring as their CTCF binding sites do not show signatures of SINE-mediated repeat expansion[434]. Mouse showed the greatest extent of SINE repeat-expansion of all the species investigated, suggesting that the process may have undergone significant acceleration during the murine-lineage evolution, with almost 4 times more SINE B2 insertions with CTCF binding sites since their last common ancestor with rats[269].

Furthermore, the dependency on DNA sequence for CTCF recruitment and its functions in insulation and long-range chromosomal remodelling could mean that

CTCF has a role to play in linking genome sequence with the evolution of chromosome higher-order organisation. This is supported by the conservation of chromatin domain structures between human and mouse reported through both linear epigenomic analysis and high-throughput chromosome conformation capture (Hi-C) comparisons[235, 435]. Rudan *et al.* further confirmed CTCF role in modulating chromosomal organisation by showing that chromosome domain structures are robustly conserved in syntenic regions, and consistent with conservation of the pattern of genomic occupancy of CTCF[312]. Conserved CTCF sites, in complex with the cohesin protein complex, display enrichment at strong TAD boundaries, with binding motif in a favourable convergent orientation. On the other hand, evolutionary divergent CTCF sites coincide with corresponding evolutionary divergent internal TAD structures[312] (Figure 1.9).

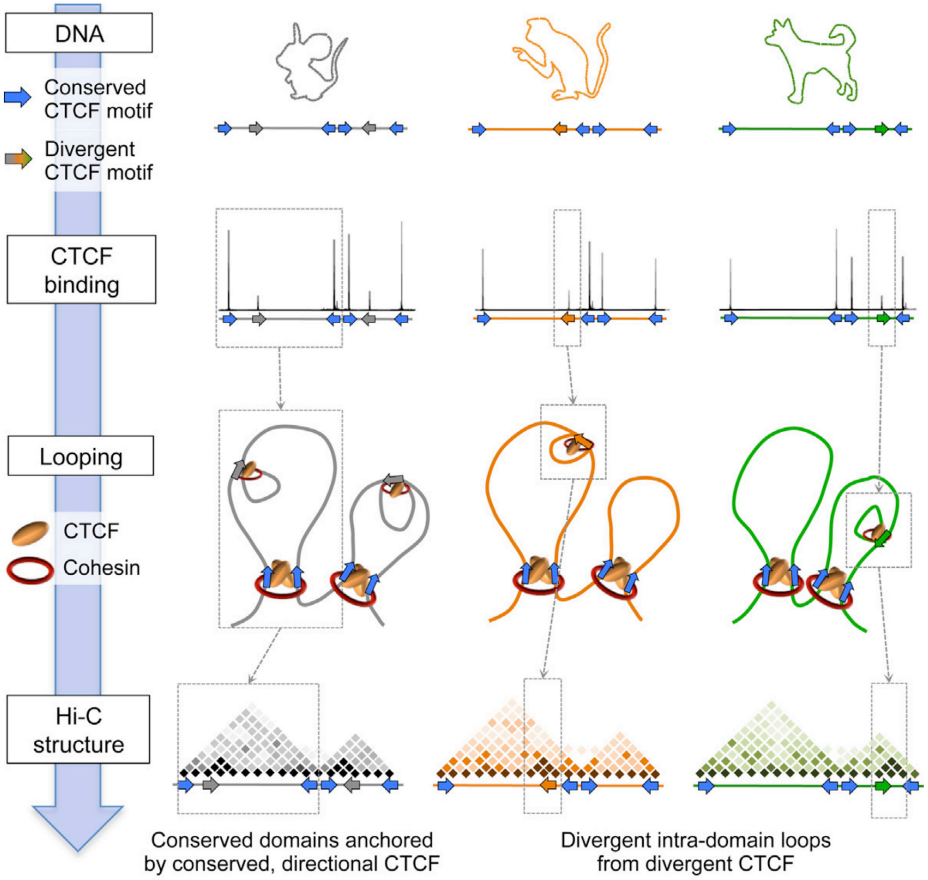


Figure 1.9: CTCF drives evolution of chromosomal domain architecture

Large-scale chromosomal domain structure (the two large grey/orange/green loops) is highly conserved across species, along with the conservation of both the CTCF binding site (blue arrows) and the orientation of its motif (direction of arrowhead) demarcating these conserved domains. Internal domain structure (smaller grey/orange/green loops inside the larger ones) is more dynamic, correlating with the evolutionary dynamics of

CTCF sites (grey/orange/green arrows) and divergence of local insulation structure (direction of grey/orange/green arrowheads). Figure adapted from Vietri-Rudan *et al.* [238].

1.3.2 Transposable elements in the evolution of gene regulation

One of the fundamental observations of the genomics era is that TEs comprise a significant proportion of vertebrate genomes[436]. 30-50% of mammalian genomes are made up of TES, mainly of the retro-transposon variety[436]. Transposable elements are repetitive sequences that have been integrated into the genomes of higher eukaryotes for millions of years[437]. Mammalian TEs can be classed into two main categories: retrotransposons, which use an RNA-mediated copy-paste mechanism to move around the genome, and DNA transposons which move directly in a cut-paste fashion instead[438] (Figure 1.10a). Retrotransposons are further classed into two main groups: long terminal repeats (LTRs), such as endogenous retrovirus (ERV)-like elements, and non-LTRs. Human LTRs are derived from ancient endogenous retroviral integration into the genome, accounting for 8% of the total length of the human genome[439] (Figure 1.10b,c). There are two subtypes of non-LTRs: autonomous short interspersed elements (SINEs) and non-autonomous long interspersed elements (LINEs). LINE-1 and Alu elements, which are also non-LTR retrotransposons, comprise about 25% of the human genome[440].

TEs contain their own promoters and regulatory elements that ensures their transcription and transposing activity in the host genome. TEs that lack this machinery such as SINEs are transposed by utilising another TEs mechanism[441-443]. The overrepresentation of TEs is an indication of their more efficient replicative capacity, which evidently surpasses that of the host genome[444, 445]. Some TEs are still active and transposing in humans, such as Long Interspersed Nuclear Elements (LINEs, mostly L1s), Long Terminal Repeat Retrotransposons (mostly ERV1-LTRs), , Short Interspersed Nuclear Elements (SINEs) of the Alu families, and SINE-VNTR-Alus (SVAs)[446]. These TEs were recently shown to be the main source of novel DNA sequences in the primate lineage, driving the evolution of novel lineage-specific regulatory elements[447]. Ancient mammalian TEs were also proposed to mediate the formation of novel gene regulatory networks in the uterus[431, 448], and have a key involvement in pluripotency[449].

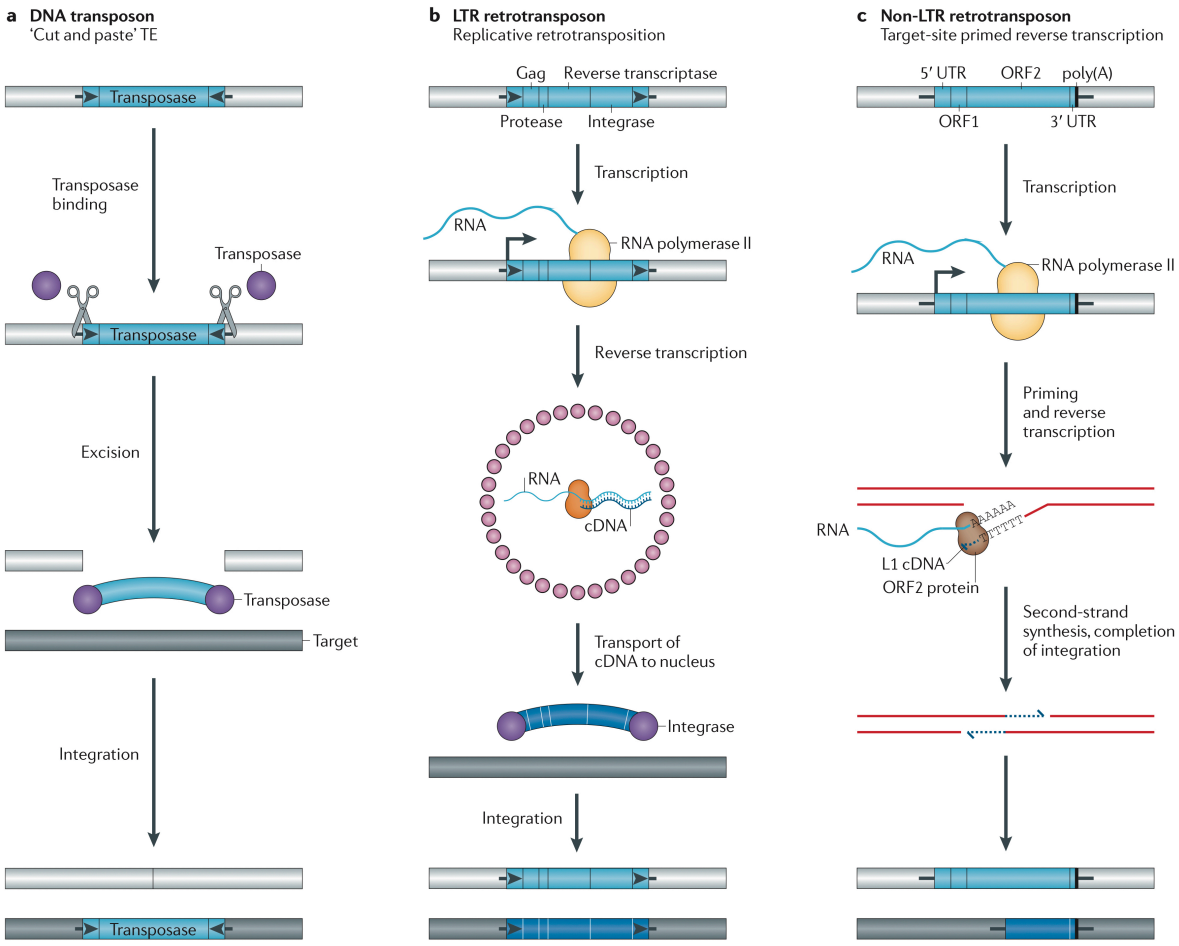


Figure 1.10: Mechanisms of TE mobilization

(a) DNA transposons are flanked by terminal inverted repeats (TIRs; black arrows), carry a transposase (purple circles), and move by a 'cut and paste' mechanism (scissors). The transposase cut the transposon from its existing genomic location (light grey bar) and pastes it into a new location (dark grey bar). (b) LTR retrotransposons are flanked by two long terminal repeats (LTRs; black arrows) and carry Gag, protease, reverse transcriptase and integrase enzymes. The 5' LTR includes a promoter recognized by the host RNAPII and transcribes the mRNA of the TE. Gag (small pink circles) assembles into virus-like particle that encodes the TE mRNA (light blue), reverse transcriptase (orange shape) and integrase. Reverse transcription copies the TE mRNA into double-stranded cDNA. Integrase (purple circles) next inserts the cDNA into the target site. (c) Non-LTR retrotransposons do not have LTRs and carry one or two open reading frames (ORFs). Transcription of non-LTRs produces a full-length mRNA (wavy, light blue line). They move by target-site-primed reverse transcription (TPRT). In TPRT, an endonuclease causes a single-stranded 'nick' in the host DNA, releasing a 3'-OH which primes the reverse transcription of the RNA.

The new element (dark blue rectangle) is 5' truncated and is retrotransposition-defective. The integration of non-LTR retrotransposons causes TSDs or small deletions at the insertion site in the host DNA. Figure adapted from Levin and Moran [450].

The first TF binding site to be found on a TE in a genome-wide scan is the tumour-suppressor gene, p53. The binding sites of p53 (~30% *in vitro*) were found on a primate-specific ERV-LTR[432]. Many other studies followed finding TF binding sites on TEs, such as the previously-mentioned CTCF and pluripotency factors (OCT4, NANOG)[269, 384, 409]. Of note is an investigation which profiled the genomic occupancy of 26 TFs in two human and mouse cell lines, and produced a quantitative estimate of TEs contribution to TFs binding sites[399, 451]. They found an average of 20% of TFs binding sites to be encoded on TEs *in vitro*[387].

TEs are demonstrably capable of rewriting existing regulatory networks in a manner consistent with the "gene-battery" model put forward by Britten and Davidson in 1969[452, 453]. The model provides a theoretical framework to elucidate how repetitive elements are an efficient mechanism for creating evolutionary divergent *cis*-regulatory modules. The nature and aspects of TEs functions make them a good fit given their inherent ability to mobilise and readily and repeatedly integrate their sequences into the genome[454] (Figure 1.11). A 2017 study was the first to show how such a module of TF binding sites in mESCs could arise in mouse-specific TEs[454]. 77% of TEs investigated showed measurable enhancer activity in mouse ESCs, and by mutating individual TF binding sites nested in the TEs, a module of TF motifs that cooperatively enhanced gene expression was discovered. The same motif module was obtained by *in silico* construction of the ancestral TE, similarly acting cooperatively to enhance gene expression. This result illustrates that TE expansion is indeed a viable mechanism to introduce novel *cis*-regulatory modules into mammalian genomes[454].

Significant association between LINEs and LTRs and lineage-specific gene family expansions was observed in both the human and mouse genomes[455]. LTRs were found enriched around the open chromatin neighbourhoods of gene families, whereas LINEs may have been involved in promoting gene duplication. The expansion of gene families, particularly in the mouse genome, seemed to have undergone two distinct phases: the first displayed an increased level of LTRs deposition, and their subsequent involvement in rewiring the gene regulatory circuits; whereas the second phase was marked by the build-up of LINEs, followed by rapid gene family duplication in a characteristically runaway process[455]. Whereas most TE classes are mainly accompanied with reduced gene expression levels, upregulation of gene expression has been reported with Alu elements in the human genome, and were the most likely of all TE classes to contribute to regulatory networks[438]. These results indicate that

young lineage- or species-specific TEs, such as SINEs, may have the biggest impact on the regulation of gene expression.

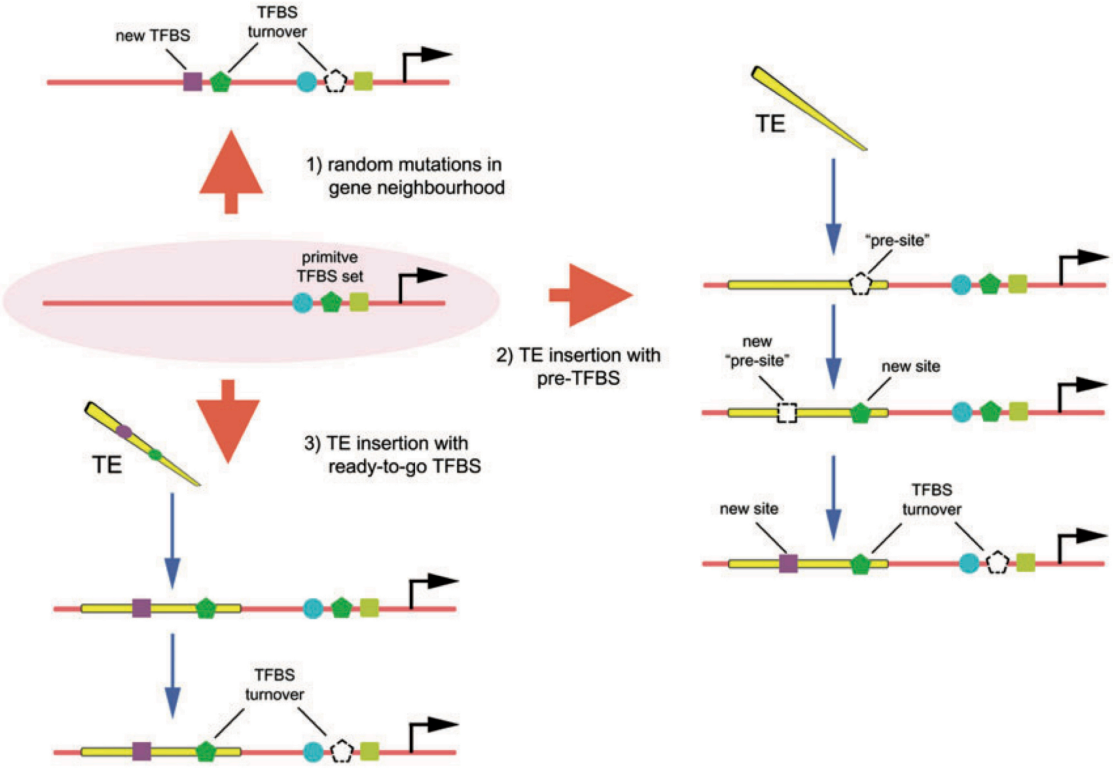


Figure 1.11: Evolution of TF binding sites via TE action

A gene (pink oval) is regulated by 3 TF binding sites (TFBS) (blue circle, green pentagon, and yellow square). 1) Novel sites could appear nearby by random mutation, causing turnover of previously present TFBS (green pentagon) or a new one arising (violet square). 2) Initial insertion of a TE in the vicinity, followed by random mutation leading to TFBS turnover and/or novel TFBS. Some pre-sites may just be one mutation from turning could eventually cause turnover of existent TFBS. Figure adapted from de Souza *et al.* [436].

The Roadmap[20] and GTEx[9] projects have shed more light on the extent of TEs contribution to gene regulation. An investigation into the association of TEs with chromatin in 24 tissues identified 112 human TE families enriched in active chromatin tissue-wide[446]. The most common TE classes were SINEs and DNA transposons, whereas LTRs were limited to particular tissues. However, genes whose expression is consistent across tissues display reduced association with TE insertions. On the other hand, different TE classes were enriched with different repressive marks: LTRs and

LINEs are enriched in regions marked by H3K9me3, while others overlapped with H3K27me3. TEs harbouring tissue-specific master regulator binding sites were enriched in tissue-specific active regulatory regions. Those included TEs in intronic enhancers and corresponded with tissue-specific variations in nearby gene expression[446].

1.4 Next-generation sequencing in regulatory genomics

Understanding the relationship between TFs and the genes they regulate, and the definition of transcription regulatory networks has been facilitated by the great advancement in the experimental methods developed in the last few decades allowing the identification of TF-DNA interactions genome-wide. Such experiments were infeasible not so long ago. While the chemical composition of DNA was identified in the 19th century, 50 years elapsed before the molecular structure of DNA was determined[456], and we had to wait 25 more years before a method to decipher its sequence was eventually developed. The original Sanger sequencing method involved a process of sequencing by synthesis (SBS) of a radioactively labelled DNA strand using the dideoxy chain termination technique[457, 458]. Sanger sequencing, or "first-generation sequencing", has later been refined and automated to use fluorescent-labelled nucleotides instead of radioactive ones, and substituted fluorescent-labelled nucleotides instead of radioactivity for gel electrophoresis[459, 460]. Further improvements came in the shape of molecular biology techniques such as recombinant DNA technology and polymerase chain reaction (PCR), producing millions of copies of sequencing fragments[461, 462]. The use of Sanger sequencing dominated biology for over two decades, crowning its reign with the publications of the full genome sequences of *D. melanogaster*, the nematode *C. elegans*, mouse and eventually human genomes[440, 463-466].

Sanger sequencing; nevertheless, has some major limitations. For example, only one DNA sequence can be analysed per lane/tube, necessitating first breaking the DNA sequence into smaller fragments, then cloning them into vectors, artificial chromosomes, transforming into bacteria, and later extraction of individual fragments from the resulting colonies[467]. During the last decade, a steady shift from automated Sanger sequencing towards newer methods referred to as next-generation sequencing (NGS). There are currently several NGS strategies available from different vendors. They all share the same basic principles of template preparation, sequencing and imaging, and genome alignment and assembly[468]. NGS technologies have expanded

into every branch of genetic and genomic research. The field in which NGS has probably made the biggest impact in is the study of the regulation of gene transcription[469].

In the following section of this chapter, I will give an overview of the NGS technology, with a focus on Solexa/Illumina sequencing that has been employed in the work of this thesis. Then, I will broadly explore the particular method of chromatin immunoprecipitation followed by sequencing (ChIP-Seq) in terms of its experimental protocol, analysis pipeline and the computational methods used downstream to interrogate its output. I will also provide an overview of the approaches to study the effects of *cis*- and *trans*-acting variants on TF occupancy and gene expression regulation. I will finish with a section on the methods used to study chromatin conformation in 3D to understand the higher-order control of gene expression.

1.4.1 High-throughput next generation sequencing (NGS)

Next-generation sequencing (NGS) is based on massively parallel sequencing or high-throughput sequencing of DNA molecules, doing away with the need for physically separating the individual reactions into separate lanes/capillaries, as is the case with Sanger sequencing. To achieve this, the sequencing reaction takes place on a solid platform, the nature of which differs depending on the particular technology, with partial spatial separation between the individual reactions. This allows billions of reactions to simultaneously occur, reducing both labour and cost enormously, whilst vastly improving the throughput of the process[467].

In this section I will focus on the Illumina/Solexa[470] NGS platform, which is incidentally the most widely-used technology for short read sequencing (Figure 1.12). Illumina uses a sequencing by synthesis technology, which had originally been developed by a company called Solexa, with improvements that increased read length and greatly improved accuracy and throughput[471, 472].

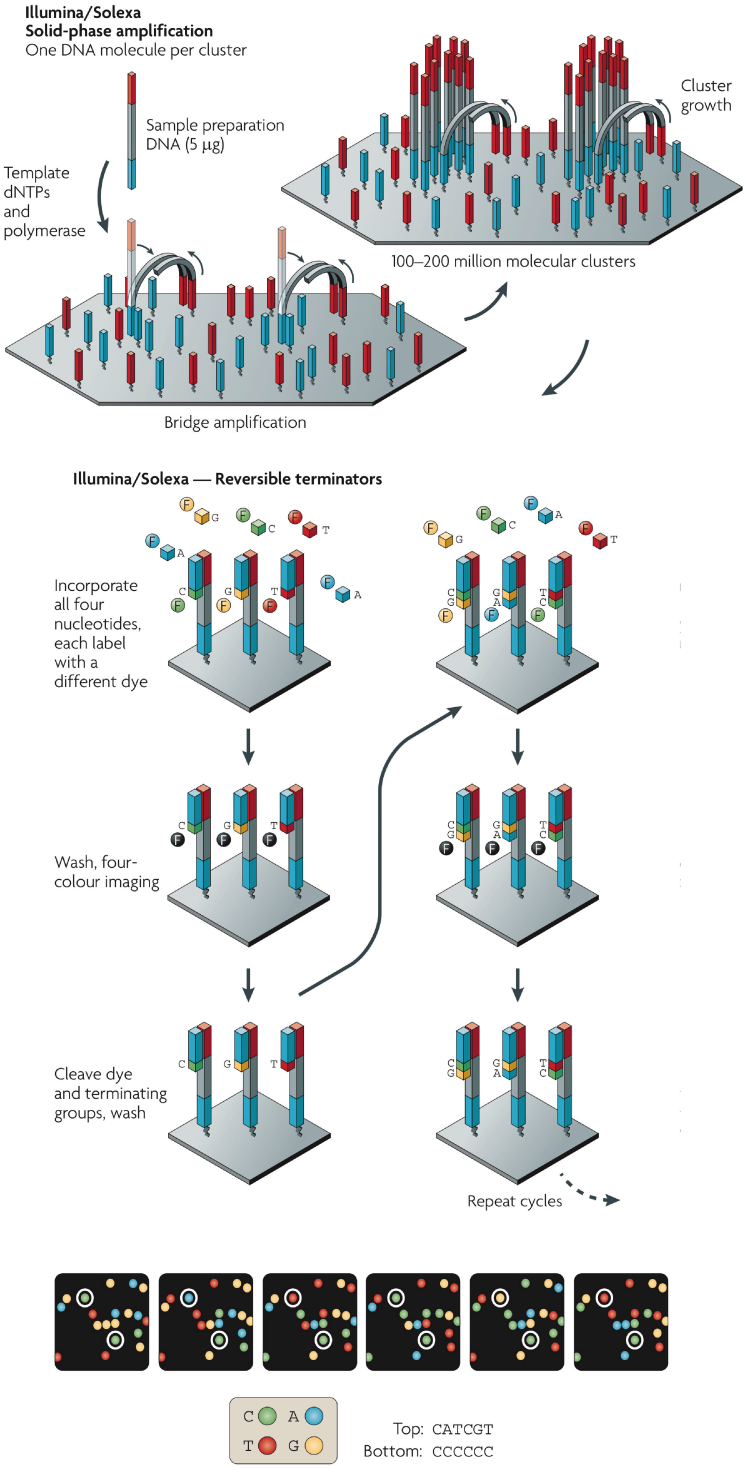


Figure 1.12: Overview of Illumina/Solexa sequencing technology.

Illumina/Solexa solid-phase amplification technology starts with priming followed by extension of the single-stranded, single-molecule template. Bridge amplification of the

fixed template with adjacent primers forms clusters. The four-colour cyclic reversible termination (CRT) method is used next. Following imaging, cleavage removes the fluorescent dyes, and cycle restarts. Figure adapted from Metzker [468].

Like all NGS technologies, Illumina offers the ability to perform sequencing of many overlapping short (50-400 bases) DNA fragments by spreading, then fixing them onto a solid platform, then replicating them in parallel[473-475]. The solid-platform amplification phase of the process used high-density forward and reverse primers covalently-linked to the glass slide. Illumina solid-phase amplification yields around 100–200 million spatially separated template clusters, with free terminal ends a universal primer can hybridise in order to initiate the sequencing reaction. Massively parallel sequencing then ensues via a DNA sequencing-by-synthesis manner, similar to Sanger sequencing, using dye terminator nucleotides[468]. This method utilises reversible nucleotide terminators in a cyclic manner that initiates with nucleotide inclusion, fluorescence imaging and cleavage. The DNA polymerase binds to the printed template and proceeds to incorporate nucleotides based on the template strand sequences until it encounters a dyed terminator nucleotide. Following the termination of synthesis on this template, the remainder of unincorporated nucleotides are removed. Imaging is then performed and the dye used for each of the four nucleotides is used to identify the incorporated nucleotide. The cleavage step removes the terminating nucleotide and the fluorescent dye, before the cycle restarts again[468]. This generates millions to billions of short stretches of DNA reads. The original template DNA molecule is thus synthesized multiple times, with each base covered by various degrees of depth. The depth of coverage could be defined as the minimum number of times each base is synthesized into an overlapping fragment. The resultant short reads are subsequently assembled, either with or without the aid of a reference genome, to recreate the original DNA template.

The combination of NGS technologies with existing biochemical methods allowed the proliferation of novel methodologies to obtain genome-wide profiles of nucleosome positioning, DNA methylation, TF binding, transcription and 3D nuclear contacts. Examples of such technologies include Transposase-Accessible Chromatin with high- throughput sequencing (ATAC-seq)[24], bisulfite treatment coupled with next-generation shotgun sequencing (BS-seq)[476], NGS of RNA molecules (RNA-seq)[477] and the various chromosome conformation assays[478]. ATAC-seq is a method for genome-wide profiling nucleosome occupancy and mapping chromatin accessibility, using hyperactive Tn5 transposase that inserts adapters into accessible regions of the genome. This allows for interrogating the accessibility of chromatin with sequencing reads, as well as identify TF binding sites and nucleosome positioning[479]. BS-seq is another powerful technique that utilises the parallel capability of NGS to

quantitatively detect the methylation status of every cytosine residue in the genome. This produces genome-wise methylation pattern of the DNA to the resolution of one bp[476]. RNA-seq facilitates the quantitative estimation of known and novel transcripts either from a cell (scRNA-seq) or a group of cells. RNA-seq provides information on gene expression level down to one bp resolution, and is useful in identifying variants in the transcriptome[477]. Capturing chromosome conformation (3C) and its scale-up variations (4C, 5C and Hi-C) are used to discern the pattern of 3D chromatin interactions. 3C is used to identify kb-scale long-range chromatin loops between to loci. 4C is used to map all regulatory interactions of a particular "bait" locus. 5C detects multiple interactions with multiple loci. The Hi-C version uses proximity-ligation, restriction-enzyme digestion and NGS to map the entire set of chromatin contacts in the nucleus[478]. Furthermore, ChIA-PET, an adaption of 3C, is used to detects interactions between particular proteins or different loci in the genome[480].

In the next section, I will explore in more detail another NGS method that has been used in this thesis for mapping the genomic location of transcription-factor binding and histone modifications *in vivo*, chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq).

1.4.2 Chromatin Immunoprecipitation followed by sequencing (ChIP-seq)

ChIP-seq is currently the standard method to detect genomic regions associated with specific proteins and TFs within their native cellular context. Robertson *et al.* first developed ChIP-seq in 2007 to allow the mapping of DNA-protein interactions *in vivo* by capturing proteins in their physiological chromatin environment[22]. ChIP-seq's popularity, compared to its predecessors ChIP-chip, DNase I hypersensitivity and array-based methods, stems from its capacity to quickly and efficiently decode millions of DNA fragments simultaneously with high throughput and relatively modest cost[481]. Using ChIP-seq, DNA-occupancy and regions of histone modification can be identified and associated with functional annotations, specific binding motifs and gene expression. ChIP-seq output can additionally be combined with other NGS applications to produce a multi-level understanding of genomic functions[482].

Sample preparation for ChIP-seq starts with chemical treatment, usually with formaldehyde, of tissue samples to cross-link proteins covalently to DNA, followed by sanitation/enzymatic digestion of the cells to fragment the DNA to the optimal size of 150-500bp. Immunoprecipitation of the target protein with its bound-DNA enriches the sample for those particular fragments relative to the starting material[483]. DNA

fragments are then sequenced as reads (typically 36–100 bp) (Figure 1.13). Most ChIP-seq analyses are performed using single-end reads, but paired-end ones are sometimes used to enhance library complexity and increase mapping efficiency at long and/or repetitive elements[484]. ChIP-seq quality is dependent on the enrichment level attained during the affinity precipitation step, which is determined by specificity of the antibody[483]. Like all biological experiments, a proper set of control samples is essential for the interpretation of ChIP-seq findings. This is particularly critical for ChIP-seq as DNA shearing during the sonication step is not uniform, and open chromatin tends to be overrepresented in the purified sample[485]. Two methods are used to provide ChIP-seq control samples: (1) "Input DNA" isolated from cross-linked cells under the same set-up as the immunoprecipitated DNA; (2) "IgG control" with an antibody that does not bind to a nuclear antigen[483].

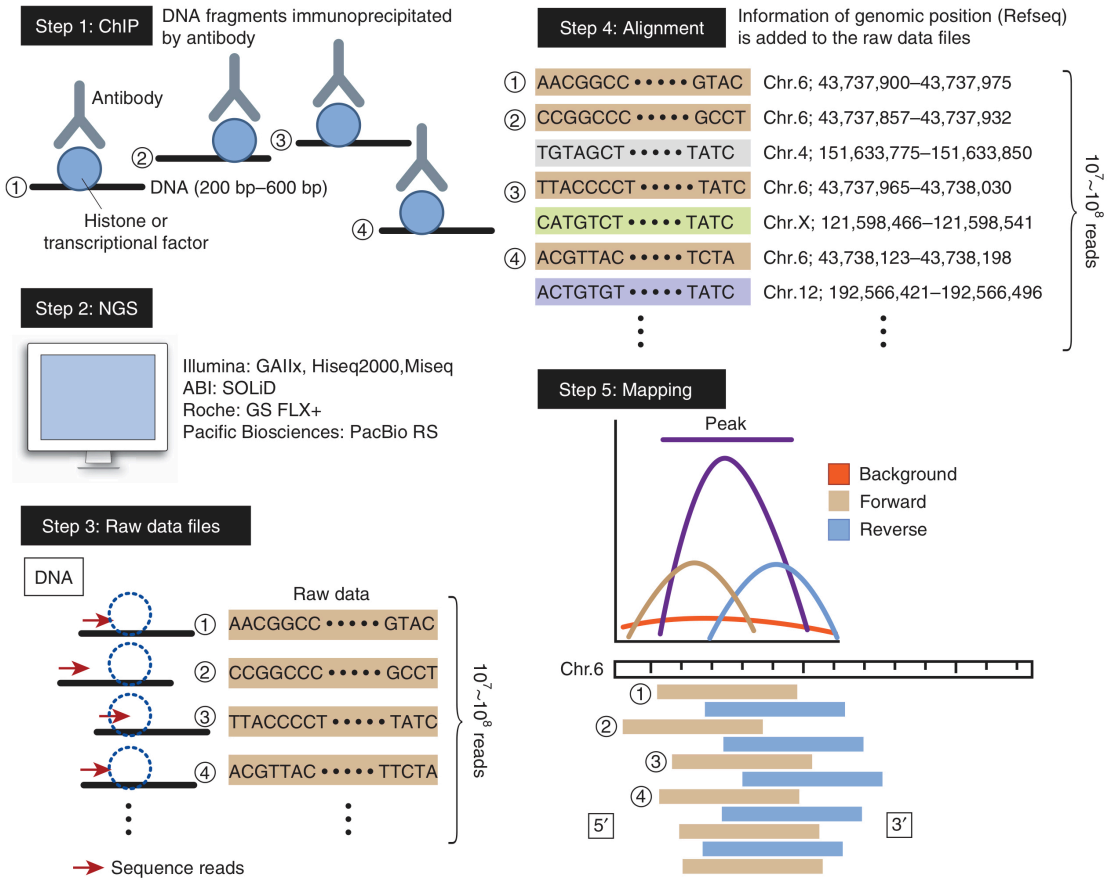


Figure 1.13: Experimental protocol for ChIP-seq

Step 1: Cells are cross-linked with paraformaldehyde and sonicated. Immunoprecipitation isolates DNA-protein fragments. The DNA fragments are between 200-600 bp long. Step 2: NGS of DNA fragments collected by ChIP. Step 3:

The NGS machine produces 10-100 million raw short reads (red arrows). Short reads are 70–100 bp long. Step 4: Short reads are aligned to a reference genome. Step 5: Peaks (purple bell-shape) are called from aligned reads using statistical models to inspect their distribution (tan, sky-blue and red distributions). Bars indicate reads from 5' (forward read, tan) and 3' (reverse read, blue) on the chromosome 6. Figure adapted from Mimura *et al.* [486].

Sequenced short reads are often outputted as FASTQ sequence format, with each nucleotide assigned a quality score (Phred-like score). The Phred score defines the probability that the base call is correct. A quality control step is generally advised to check on the read length distribution, GC content, quality scores, overrepresented sequences and k-mer content[487]. FastQC[488] is a common toolkit for this purpose, and other tools are also available for further modifications to the data including adapter sequence removal and 3'-end sequence trimming.

Sufficient sequencing depth is essential for effective analysis of ChIP-seq data, as the number of the DNA-protein regions identified positively correlates with the sequencing depth[489]. Sites with weaker binding achieve statistically significant enrichment over the background only with a greater number of reads, and without adequate sequencing depth, valuable information about their occupancy and functionality may be lost. The level of sequencing depth is based on the number and size of the protein/histone marks binding sites and the particular species genome size. Based on ENCODE guidelines, for mammalian TFs and histone modifications a coverage of 20 million reads is usually sufficient to capture the thousands of binding sites of those factors, which tend to occupy localised, narrow sites[483]. Factors with substantially more binding sites, such as the ubiquitous RNAP II, or broad histone modifications, need to be sequenced significantly deeper, up to 60 million reads, in order to properly capture a reasonable genome-wide occupancy pattern[484]. Furthermore, control samples must always be sequenced much deeper than factor-bound samples in ChIP-seq in order to ensure proper coverage of the genome and non-repetitive autosomal DNA[490].

Short read sequences are subsequently mapped onto the genome using alignment tools. ChIP-seq analyses do not usually require aligners to handle gapped alignments (INDELs) as the sequencing reads should normally not harbour any, except in the case of cross-species comparative studies, which sometimes map reads onto another species' genomes. If the sample contains important information about single nucleotide polymorphisms (SNPs) and INDELs, such as the case of heterozygous variant analysis, an allele-specific mapper needs to be used[491-493]. Widely-used alignment algorithms are coded to handle ambiguities of repetitive sequence and

sequencing errors to a reasonable degree. The most common alignment algorithms are based on the Burrows-Wheeler Transform and designed for short read alignments. Long read alignment is usually carried out using hash table-based methods and the Smith-Waterman alignment[487]. Most-commonly applied aligners include Burrows-Wheeler Aligner (BWA)[494], Bowtie[495], MAQ[496], SOAP2[497], GSNAP[498] (which is specifically coded to handle mismatches in heterozygous sites), and great many others. A recurrent issue of ChIP-seq alignment is whether to include reads mapping to multiple locations on the reference genome. Opting to include multiple mapped reads may substantially increase the number of reads available for downstream peak-calling and analysis, thus enhance detection statistics[492]. This; however, also increases the false discovery rate (FDR). Except in the case of in-repeat analysis, uniquely mapped reads of a particular TF should suffice[499].

Following the mapping step, the ChIP-seq experiment signal-to-noise ratio (SNR) needs to be determined using quality metrics such as strand cross-correlation or IP enrichment estimation[490]. These metrics are designed to identify the various ways a ChIP-seq experiment may fail, such as insufficient enrichment by the IP, inadequate fragment-size selection, or poor sequencing depth. The strand cross-correlation metric is now already built-in some of the more common peak-callers such as MACS (version 2) and SPP. ChIP-seq short reads are now aligned, tagged with metadata and alignment scores, and ready to be used for peak-calling of DNA-protein regions and all further downstream analysis.

1.4.3 Peak-calling and downstream computational approaches

Peak-calling software are specialised packages designed to identify TF or histone mark genomic occupancy based on ChIP enrichment. The peak calls are outputted in the form of a ranked-list of regions based on either they read signal or p-value computed based on the significance of enrichment[483]. The most widely-used peak calling packages include MACS[500], PeakSeq[501], and SPP[502], but there many more[503]. Peak-calling programmes use the control sample to estimate the background distribution of the regions identified in the TF samples. Early peak-callers utilised the Poisson model that assumes a uniform genome-wide background distribution for the control reads. Due to the observed enormous variation in the read distribution that cannot be accommodated into the Poisson model, the model was extended into a negative binomial model adopting an overdispersion approximation to account for this variation. Later peak-callers extended it further to a zero-inflated negative binomial model because of the zero-inflated read distribution in zero- or low-mappable

regions[482]. The work presented in this thesis was analysed using MACS, a peak-caller that adopts a local Poisson model[500].

Model-based Analysis of ChIP-seq (MACS) scales the total control read (or tag) count linearly to equal the total tag count[500]. Due to inherent biases in the sequencing method (5' to 3' sequencing direction), more tags accumulate in the forward part of the peak region than in the reverse one, resulting in a peak shift (Figure 1.14). MACS first estimates the distance of the peak shift from both DNA strands, and accordingly shifts the tag distribution to properly cover the true TF-binding site.

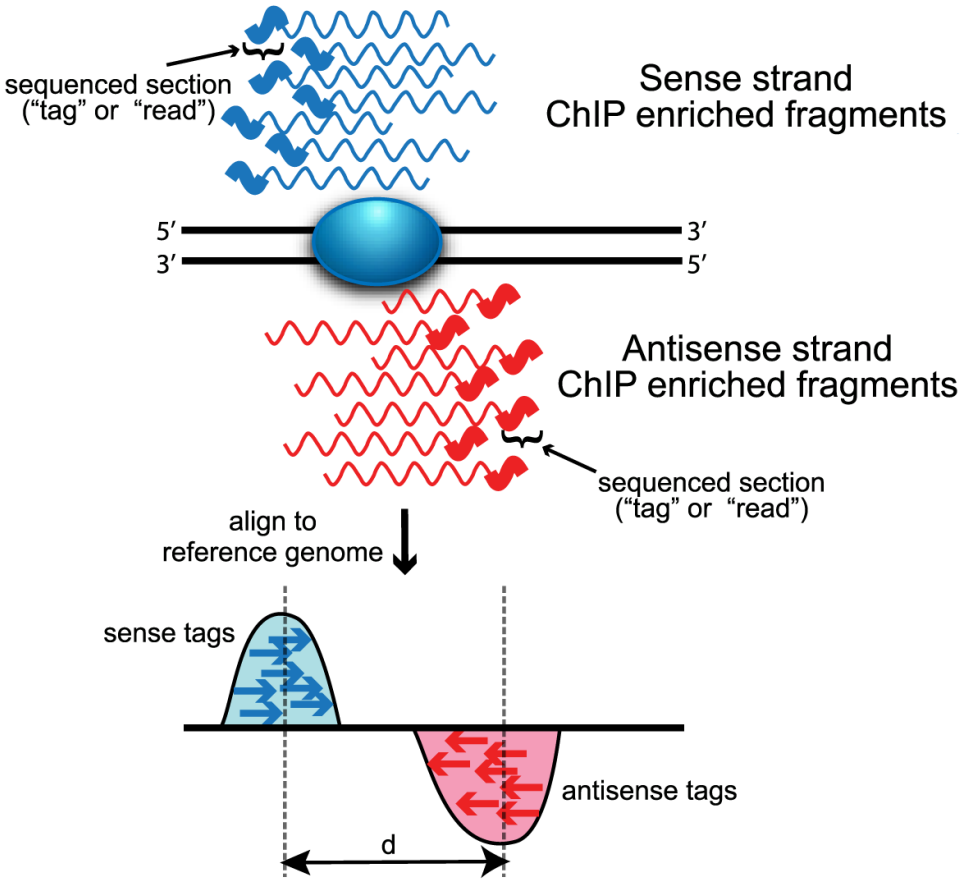


Figure 1.14: Stranded bias in tag density of ChIP-seq experiments.

The blue-shaded oval represents the TF bound to DNA (black lines). Wavy lines denote sense (blue) or antisense (red) short reads from ChIP-seq experiment. The thicker end of the line indicates the short-read tag. Following alignment to a reference genome and being given chromosomal coordinates (red and blue arrows), the distance between peaks (d) corresponds to the average sequenced fragment length. Figure adapted from Wilbanks and Facciotti [504].

The genome-wide tag distribution is then modelled with a Poisson distribution, allowing the parameter λ_{BG} to capture the mean and the variance, and MACS calculates a p-value for each potential peak. MACS shifts every tag by distance(d)/2, utilising a 2d sliding window along the genome to identify significantly-enriched candidate peaks (Figure 1.14). MACS merges overlapping candidate peaks, extending each tag a d number of bases to the centre of the region. The summit, the location of TF binding, is the position where the highest fragment pileup is observed. MACS dynamically estimates the local Poisson distribution for the surrounding region of the current window at a distance of 1, 5, and 10 kb from the centre of the window and computes the Poisson parameter for the window using the formula:

$$\lambda_{local} = max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$$

This approach ensures small peaks in low-enrichment loci are detected and avoids local genome biases. Finally, using the same parameter MACS conducts a sample swap to find ChIP peaks over control and control peaks over ChIP. This step empirically computes the FDR as (No. Control Peaks/No. ChIP peaks)[500].

The TF binding site is usually in the range from 8-20 bps[505]. Peak regions called from ChIP-seq data, on the other hand, are always much longer, in the range of several hundreds of bps long. Motif discovery is subsequently essential to pin down the position and identity of the actual binding sites within these peaks. The general principle behind Motif discovery is that the regions of the ChIP peaks should in all or most of the cases carry an oligonucleotide sequence that is sufficiently similar to be identified as the binding motif of that particular TF. This oligonucleotide sequence could be the same previously reported motif of its TF, or a newly discovered one. The same oligonucleotide should also not have any likeness to a random set of sequences built at random with a generator of "biologically feasible" DNA sequences, or appear at the same frequency[506]. This enriched oligonucleotide sequence is then used to identify the actual motif to which the TF binds based on PWMs created for each base in the sequence. Motif analysis can additionally be used to detect added motifs in the neighbourhood of these regions which could serve as binding sites for other TFs functioning as binding partners in cooperative manner, and thus involved in creating regulatory modules[507]. The MEME Suite of tools offers a wide variety of motif analysis options, ranging from discovery of *de novo* motifs to searching a reference motif database (such as JASPAR, UniPROBE, TRANSFAC, etc.)[508, 509]. JASPAR Vertebrates and UniPROBE Mouse is the one mostly used for vertebrates, and was the one utilised in this thesis whenever motif discovery or enrichment was required[510].

Following peak-calling, the ChIP-seq peaks obtained need to be functionally associated with genomic regions where they can undertake their roles, including

promoters, transcription start sites and intergenic regions. Packages such as BEDTools[511, 512] provide a universal toolbox for the systematic analysis of peak calls. BEDTools can be used to measure the distance from each peak to the nearest genomic feature of interest, or to up/downstream genes, and combine the findings with results from other NGS methods such as expression data from RNA-seq, methylation information from BS-seq, chromatin interactions from Hi-C.

This is, however, easier said than done. associating distal TF binding sites to their true target genes is not always feasible. For example, NF-γ, which is strongly associated with promoter and TSSs, has most of its binding sites located a long way from those regions[513]. Restricting the analysis to TF bound near promoter sites only is not an option either, as this misses all distal targets for the TF, and severely limits understanding the extent of its regulatory potential. Furthermore, the number of TF binding sites identified in a ChIP-seq experiment is high enough that if each one was associated with its closest gene, a significant portion of the genome will be covered as targets[514]. Tools exist which attempts to tackle the problem such as GREAT and HOMER[515, 516]. However, basic criteria need to be applied first to reduce the number of TF considered for further downstream analysis, such as setting threshold on the distance from the TSS, and read enrichment signal. For instance, the Roadmap Epigenomics project an enhancer was only associated with a target gene if its TSS was located at less than 30 kb away[20].

1.4.4 Methods of studying sequence variation effects on gene regulation

Adapting ChIP-seq and other NGS technologies to address the potential effects of genetic sequence changes on the presence or absence of regulatory event of interest such as TF occupancy has led to an increased understanding on how variants within and across populations/species shape the regulatory landscape. DNA sequence changes within regulatory elements have the capacity to influence TF binding stability and its ability to induce its effects on transcription or modifying chromatin state, ultimately affecting the regulatory potential of the region and its impact on transcriptional regulation.

To study the regulatory effects of sequence variants (see 1.1.4 for more details), two methods have been used extensively: expression quantitative trait loci (eQTL) mapping (using gene expression levels as a quantitative phenotypic trait), and allele-specific expression divergence between parental strain and their F1 hybrid in genetically inbred organisms. eQTL are polymorphic DNA variants that are associated with changes in gene expression and phenotypes[517]. Modern eQTL studies use RNA-

seq to provide allele-specific gene expression levels[518]. RNA-seq produces 10s-100s millions of sequence tags that provides a complete profile of gene expression and the isoform structure of each gene[519]. eQTL mapping have been utilised to identify regulatory regions driving variation in mRNA levels and differ between local regulators acting in short genomic range in allele-specific fashion (*cis*) and distant-acting regulators (*trans*)which influence the transcriptional processes by affecting the availability of other factors involved in gene expression, resulting in similar expression levels from both alleles[252, 518].

eQTL analysis comprises four main steps: DNA genotypes processing, RNA-seq tags processing, counting of total sequence reads and eQTL mapping[520]. First, DNA reads are mapped back to their reference genome, the genotypes are called and their haplotypes are imputed using a phasing algorithm. Next, RNA-seq reads are aligned to the same reference and/or the two haploid genomes imputed based on the results of the phasing programme[521]. After that, total read counts per gene, per sample, as well as the allele-specific reads per allele of a gene, per sample are counted, removing reads with low mapability and quality scores. eQTL mapping follows whereby variation in allele-specific expression and total gene expression are associated with a *cis*/*trans* variants using a beta-binomial distribution to test for similarity/difference in gene expression between the two alleles of a gene[520]. A hierarchical Bayesian model has been suggested to test the disparity of gene expression across alleles, combining information across genome-wide loci[522].

A variant on the method involves chromatin immunoprecipitation quantitative trait loci (ChIP-QTL)[523], which combines this approach with identifying TF-DNA contacts as discussed in the previous sections. eQTL has been implemented on a genome-wide scale successfully in a number of studies to investigate distant acting variation, epistatic interactions, and determining gene expression divergence phenotypes[524-527].

Resolving the regulatory effects between *cis*- and *trans*-acting variation in eQTL studies remains fairly challenging despite advances in both experimental techniques and computational approaches[528, 529]. First, eQTL analyses require a vast number of genetically diverse samples to reach sufficient statistical power for detection[530-533]. Furthermore, eQTL analyses cannot fully distinguish between *cis*- and *trans*-acting elements. Some of the *trans*-acting variants may be located in close proximity on the same DNA molecule of the target gene, and some *cis*-acting variants may be distantly located[534, 535]. In addition, *trans*-eQTL have much smaller effect sizes, are less robust, less common and require a high number of association tests to investigate than *cis*-eQTL[536], which in turn reduces the statistical power, hindering

their detection[537]. Additionally, *trans*-eQTL suffer from the same confounding factors that influence their *cis* counterparts, be they biological (e.g. haplotype effects, tagging *cis*-eQTL), technical (probe binding sites variation) and statistical (missing genotypes, population structures)[538].

The F1 hybrid method, in contrast, can avoid many of those caveats, and reliably resolve the regulatory changes brought about by *cis*- and *trans*-acting variation on gene expression and TF occupancy. In this approach, variation between the two parents allow allele- specific expression to be evaluated. In F1 hybrid of the two F0 parental strains, *cis*-acting regulatory variants appear linked to their target gene reflected in allele-specific expression. *Trans*-acting regulatory variants affect both F0 alleles equally due to the shared nuclear environment. These two fundamentally different effects allow comparison of differential expression between the F0 strains and the allele-specific expression in the F1 hybrids, resolving the regulatory divergence in *cis* and *trans* across the entire transcriptome. Genes differentially expressed due to one or more regulatory variants acting in *cis* result in a ratio of allele-specific expression in F1 hybrids equal to the ratio of expression between the parent strains. On the other hand, if both alleles are expressed equally in the F1 hybrids, the difference is due to one or more *trans*-acting regulatory variants[260]. Whereas eQTL studies require a large number of crosses/samples, it is a major additional advantage that the F1 hybrid method requires only two parental strains and their F1 hybrid for analysis[252]. This approach has been used to study allele-specific gene expression in F1 hybrids in yeast[539-541], fruit flies[542, 543], and mice[257, 260, 544, 545].

To apply this method to polymorphic sites that are linked to regulatory variation, ChIP-seq peaks can be investigated to search for sequences that align across a heterozygous base in F1 hybrids. An observed difference in the binding intensity signal of one allele versus the other suggest a possible allelic effect on TF binding. For a TF binding site with two alleles, the binding signal from both alleles in F0 and F1 would be equal if no sequence effect is present, resulting in non-differential binding. On the other hand, if the binding signal differs between F0 strains or the F1 hybrids, this indicates sequence-specific effects on the regulation of TF occupancy. These could be due to *cis*- and/or *trans*-acting variation[493]. This type of analysis requires adapting the standard ChIP-seq analysis pipeline discussed above to accommodate sequence variants that may introduce bias at the alignment step as heterozygous sites where reads identical to reference genome are aligned at a higher rate due to 'mismatch' penalty imposed on the non-reference allele. In the F1 hybrid analysis, two reference genomes may be created, each containing one allele for the variant site, making it possible to combine the separate alignments of reads to each of these genomes. Alternatively, one can map F1 reads to both reference genomes of the F0

strains, and subsequently combine aligned reads for further analysis[257]. Furthermore, there are allele-aware aligners that dynamically account for multiple alleles during alignments, such as the Genomic Short-read Nucleotide Alignment Program (GSNAP)[498]. In sum, this type of analysis requires particular care and consideration for alignment of sequence variants in order to allow the accurate detection of differential binding signals from TF binding sites.

1.4.5 Methods of studying genome folding effects on gene regulation

Resolving the dynamic conformation of the folded genome within the nucleus is paramount to our ability to decipher the role of genome topology in the regulation of gene expression. Two major approaches are currently taken to study the 3D structure of the genome: chromosome imaging (e.g. fluorescence in situ hybridization of DNA or DNA-FISH), and chromosome conformation capture (3C), of which Hi-C (high-throughput chromosome conformation capture) is the most prominent method[546].

DNA-FISH contributed to the field of studying genome folding by allowing the visualisation of nuclear DNA topology in space[547]. This approach, on the other hand, is limited by low throughput that constraint the number of genomic interactions that can be investigated for each run. However, the concept of utilising matrices of contact frequencies to deduce chromatin folding[548] revolutionised the field, and resulted in the development of an array of high throughput 3C-based assays including 4C, which selects for interactions of one region with the genome ('one versus all')[338, 549], 5C, which enriches for contacts of a larger genomic stretch at high resolution ('many versus many')[550] and ChIA-PET, which combines ChIP, 3C proximity ligation and Paired-End Tags sequencing to characterise genome-wide chromatin interactions[551].

Hi-C is a genome-wide 3C method that is used to map chromatin contacts genome-wide at a scale of few hundreds of kilobases to megabases ('all versus all')[232]. In this method, DNA restriction fragments are crosslinked with formaldehyde, labelled with biotin and subsequently ligated. The biotin labels are then removed from unligated fragments by the exonuclease activity of T4 DNA polymerase, leaving ligated fragments enriched with the biotin label in the sequencing library[232, 552]. Sequenced reads aligned to the reference genome are used to assemble the Hi-C dataset, and interacting DNA fragments are binned at a range of resolutions (from 5–100 kb bins), with bin sizes depending primarily on the depth of the sequencing library[553]. Sequencing depths of 200–400 million reads are required for conducting standard Hi-C in mammalian genomes[546]. Binning reads produces a symmetric matrix of bins (genomic loci) for each row and column in the matrix. Matrices for the whole genome,

selected chromosomes, or genomic regions of interest are generally displayed as heatmaps that visualise DNA-DNA interactions by the number of reads each bin contains. There exists a host of computational pipelines for Hi-C data processing, such as HOMER[516], Juicer[554], HiCUP[555] and HiCPro[556].

3C-based methods also do have their limitations. Low efficiency of ligation and the genomic topology of the two DNA interacting DNA fragments are known to affect the detection of such contacts[546]. Furthermore, chromatin interactions detection is dependent on the fragmentation step, and 3C-based methods in general tend to be biased towards the detection of low- vs. high-order chromatin contacts, resulting in the underestimation of the impact and functional implication of the more complex interactions in the regulation of gene expression during chromatin folding[546].

1.5 Thesis Outline

In this introductory chapter, I presented the biological background, with particular emphasis on the evolution and functional roles of CTCF, underlying the work carried out in this thesis. Furthermore, I provided an overview of the methods relevant to the research projects conducted here, including the experimental approaches utilised to generate the data, and the subsequent computational analyses.

Chapter 2 presents the findings of an investigation into the impact of novel subspecies-specific CTCF binding sites in two *Mus* genus mouse subspecies, *Mus musculus domesticus* and *Mus musculus castaneus* that diverged one million years ago. This chapter focuses on CTCF occupancy difference between sites conserved in binding in the two subspecies and the evolutionary young sites bound in one but not the other subspecies (Figure 1.15). The analysis revealed a recent expansion of SINE B2-B4 transposable elements that resulted in the creation of novel subspecies-specific sites. A subset of evolutionarily young sites exhibited conservation of binding across multiple tissues in *M. musculus domesticus* (BL6). We additionally found a tandem duplication of a regulatory region comprised of a CTCF binding site and an interferon gene, forming a 15-gene BL6 specific immune locus.

Chapter 3 uses an F1 hybrid system to look at the divergence of CTCF binding influenced by regulatory variation between the two closely related mice. This chapter investigates a set of CTCF sites whose binding site between the two species is characterised by the presence of single nucleotide variants that can be used to differentiate allelic-specific binding, and the regulatory variation in the binding site (Figure 1.15). Whilst *cis*-acting regulatory variation is the most common, we observed pervasive *trans* effects in allelic-specific binding. We also investigated the tissue-wide

conservation of occupancy of these sites and described the pattern of evolution of lineage-specific sites, and the mode of inheritance CTCF binding sites demonstrate.

Chapter 4 presents the analysis of the functional regulatory potential of CTCF binding in the context of both sets of CTCF sites derived from Chapter 2 and 3. Namely, this chapter compares the functional and regulatory characteristics of CTCF binding in binding sites either on the basis of their evolutionary and tissue-specificity (Chapter 2) and or the binding variation they display in the form of *cis*/*trans* regulatory variants (Chapter 3). We investigated the enrichment in repeat elements in *cis*/*trans*-acting variants and whether their SINE-derived sites stem from the recent, subspecies-specific expansion (Chapter 2) or is older. We also analysed the differences observed among the various types of CTCF binding sites in terms of their binding in active regulatory elements, TAD-boundary association and their interaction with cohesin-complex proteins.

Chapter 5 provides general conclusions of our findings, and presents avenues for future research in which the results in this thesis could be taken.

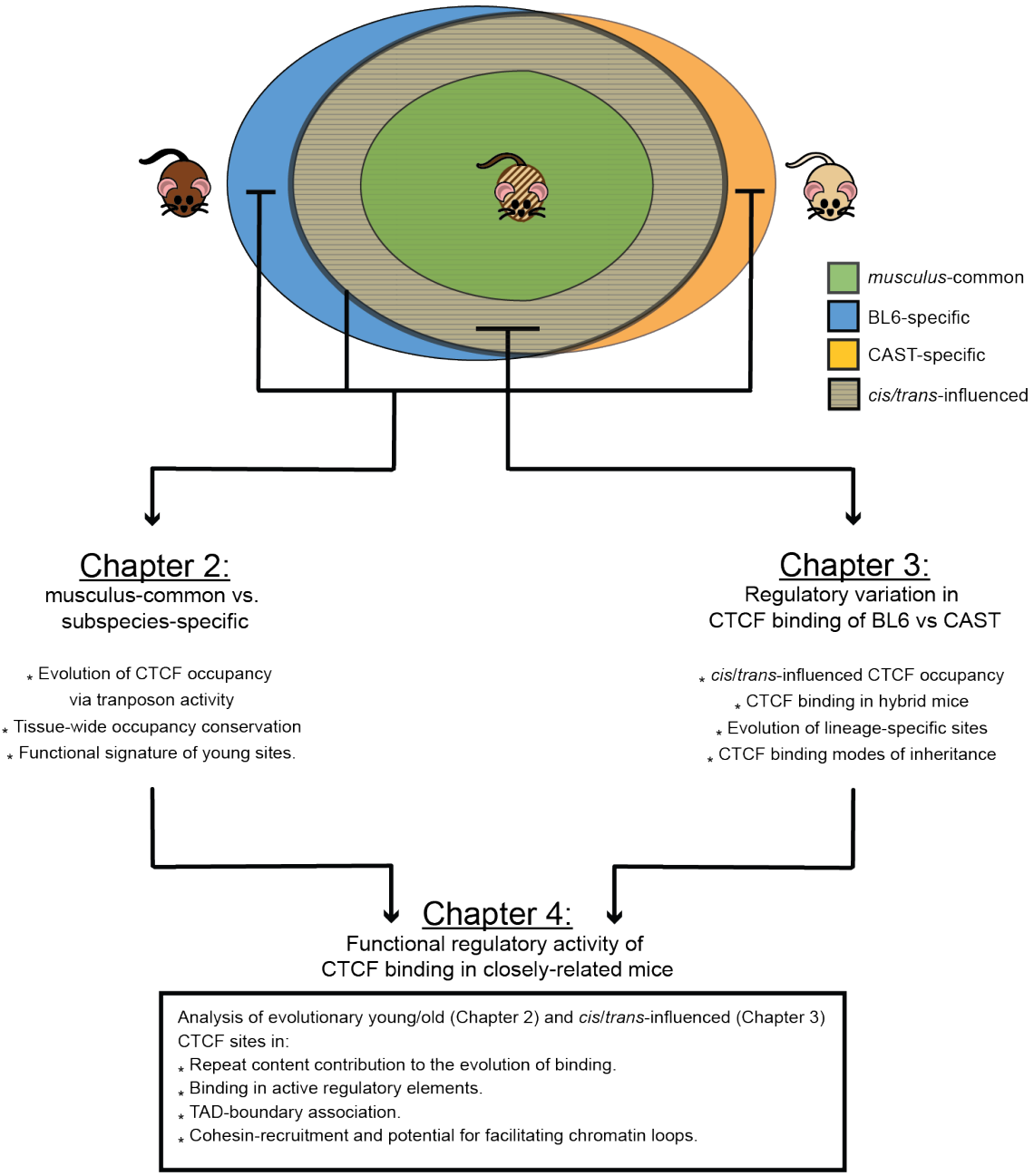


Figure 1.15: Schematic diagram of the thesis structure and the analysis pipeline.

Chapter 2

Functional signatures of evolutionarily young CTCF binding sites

2.1 Introduction

The regulatory regions of the genome predominantly lie in the non-coding sequences of the genome, and differences in those regions are thought to drive the distinction in gene expression across tissues and cell-types[401, 413, 414, 557]. Whilst protein-coding genes have been subject to strong selective pressures, as revealed by interspecies comparisons of mammalian genomes[378], tissue-specific transcription factor binding frequently diverges [6, 260, 269, 362, 392, 394, 408, 409, 558]. Changes in the regulatory non-coding genome seem to be the major force behind the variation in transcription factor binding between closely related species with both small-scale sequence variation[257, 393] and novel species-specific repeats playing important roles[559].

Repeat elements have been shown to drive some of the changes in transcription factor binding, and the regulation of gene expression by altering the non-coding genome[387, 560-562]. In particular, CTCF binding has been shown to be the product of waves of species-specific expansion of repeats across several mammalian lineages that carried its canonical motif into novel genomic locations[269, 384]. This repeat-driven, TF binding site birth mechanism has been observed in other cases of tissue-specific transcription factors in stem cells[409] and in pregnancy associated tissues[431], suggesting that repeat expansions are a common mechanism used to remodel mammalian genomes[387].

CTCF binding motifs have undergone expansion in Murine lineages via a rodent-specific family of transposable elements[269, 384]: the B2 short interspersed elements (SINE-B2-B4)[563]. Although the evolution of CTCF binding in the *Mus* genus via this mechanism has been investigated recently[559], the potential functional roles of novel, species-specific CTCF binding sites and the pattern of their genomic occupancy is not yet known.

Leveraging the high-quality genome sequences of different mouse strains and species within the *Mus* genus created by the Mouse Genomes Project[465, 564-566], we illustrate how highly active, expanding repeats have remodelled CTCF binding, and thus chromatin and transcription, in two *Mus* genus subspecies sharing a common ancestor one million years ago (MYA): *Mus musculus domesticus* (C57BL/6J *or* BL6) and *Mus musculus castaneus* (CAST) (Figure 2.1a). Subspecies-specific binding of CTCF reveals signatures of function, genomic occupancy patterns, and tissue-independent characteristics that are largely similar to CTCF sites common between the subspecies. More importantly, a subset of these subspecies-specific sites are bound in multiple tissues, suggesting active participation in loop formation and the creation of novel regulatory modules. We also found a cluster of interferon genes with subspecies-specific CTCF binding sites on mouse chromosome 4 with a high degree of sequence similarity that apparently arose via a recent, BL6-only, tandem duplication event. Taken together, these results demonstrate the evolutionary pace at which CTCF binding sites expand in the genome and acquire functionality.

This investigation is the result of a collaboration between Dr. Paul Flicek's research group at the EMBL European Bioinformatics Institute and Dr. Duncan Odom's laboratory at the Cancer Research UK Cambridge Institute. Dr. Christine Feig performed all of the wet lab experiments for this project. Dr. Jonathan M. Mudge did the manual genome annotation. I carried out the computational analysis, except where otherwise specified.

2.2 Methods

2.2.1 Experimental methods

2.2.1.1 Animal breeding and sample collection

The experiments were conducted using mice from two strains: C57BL/6J (stock number: 000664, source: Charles River Labs) and CAST/EiJ (stock number: 000928,

source: The Jackson Laboratory). All biological replicates collected for the purposes of this investigation were sampled from adult male mice, 8-12 weeks of age, and harvested between 8 and 11 a.m. All animals were kept in similar husbandry conditions in the Biological Resources Unit of the Cancer Research UK–Cambridge Institute under a Home Office Licence.

Sampling of liver by perfusion was done on mice post-mortem, followed by tissue dissection. Harvested tissue samples were quickly chopped and transferred into a cross-linking solution with 1% formaldehyde in preparation for ChIP-seq protocol. Tissue samples were incubated for 20 minutes before quenching with 1/20th volume of 2.5 M glycine, then for a further 10 min. Samples were subsequently washed with PBS, flash-frozen and stored at −80 °C.

2.2.1.2 Generation of CTCF ChIP-seq Data

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments for the CTCF protein[567] were performed using liver tissue sample isolates from two closely-related *Mus* subspecies: *Mus musculus domesticus* (C57BL/6J *or* BL6) and *Mus musculus castaneus* (CAST). The libraries were sequenced at 100 bp in paired-end fashion. In each case, three biological replicates from different 8-week old male mice, and a matched control liver sample from another animal. Technical details of the ChIP-seq procedures and antibodies used were previously reported[559].

2.2.2 Computational methods

2.2.2.1 Sequence Alignment and Peak Calling

All libraries were retrieved as raw ChIP-seq FASTQ reads were subject to quality control using standard parameters in FastQC version 0.11.5[488]. Good quality reads (min Phred score >= 30) were subsequently aligned to most recently available genome assembly in Ensembl (GRCm38 for BL6 and CAST_EiJ de novo assembly at ftp://ftp-mouse.sanger.ac.uk/, later in the Ensembl release 84 for CAST). We aligned the sequence reads to the reference genomes using BWA version 0.7.12[494] using both paired ends reads for each biological replicate and control. Aligned reads were afterwards filtered for duplicate and non-unique reads, sorted and indexed using SAMtools version 1.2[568]. CTCF binding sites were identified by peak calling from aligned sequence reads using MACS version 2.1.0[500] with a p-value threshold of 0.001 to call peaks representing CTCF binding regions. Peaks found in at least two biological replicates out of the three were used for downstream analysis. Motif analysis focused on the summit point (±50 bp) of each identified CTCF binding sites using the MEME

suite version 4.10.2[508, 509]. Most common motif found in each dataset is reported in **Figure 2.1b**.

2.2.2.2 Interspecies comparisons

We quantified the conservation of CTCF occupancy in one of the two mice subspecies using the orthologous alignments of the CTCF binding sites in the other subspecies. Interspecies comparison between BL6 and CAST was performed first using a multiple alignment of 15 de novo assemblies of laboratory and wild-derived strains genomes within *Mus musculus*[566, 569]. Orthologous regions with a CTCF binding site present in both subspecies was considered a "*musculus*-common" site, whilst sites found in only one of the subspecies, but absent from the other, was considered "subspecies-specific".

2.2.2.3 Repeat Masking of CTCF binding sites

CTCF binding regions from *musculus*-common and subspecies-specific sets of the data of both subspecies were screened for repeat elements using RepeatMasker 4.0.5[570] (http://repeatmasker.org) using the rodent repeat libraries from RepBase (v20140131) for the two murine subspecies, with the cross_match search engine, running in slow speed/sensitivity, masking for interspersed and simple repeats and RepeatMasker matrix choice for GC level. Fragmented hits found to be part of the same repeat were merged as one.

To calculate the background representation of the 4 superfamilies of transposable elements (TEs) (SINEs, LINEs, LTRs, DNA transposons) in the mouse genome for comparison with *musculus*-common and BL6 subspecies-specific sites enrichment, the sum total of the sequences occupied by each TE superfamily divided by the total length of the genome. We retrieved the full set of TEs for the C57BL/6J mouse genome from those published in Thybert et al[559]. To derive the random set of genomic sequences, we used the BEDTools version 2.2.5.0[511, 512] shuffle tool to generate sequences equal in number and length to the total number of CTCF peaks obtained from our ChIP-seq libraries. Random sequences were matched for the chromosomes, but non-overlapping with any of the sequences in the CTCF peaks set.

We used the intersection between CTCF peaks and the full set of the four TE superfamilies to derive the proportion of sequence occupied in each CTCF binding site and the relative age of the repeat element present. To determine the fraction of sequence occupied, we used BEDTools intersect 2.2.5.0 with the option -wo to return the overlap between the peak sequence and the repeat, then divided the overlap by

the total length of the peak to obtain the percentage of sequence occupied by TE for every single peak/random sequence. We defined the relative age of a repeat element as the percentage of sequence substitutions in each repeat from the consensus. The higher the percentage of substitutions in TEs compared to the consensus, the older the sequence is.

2.2.2.4 Repeat Content Analysis of liver-specific transcription factors (TF) binding sites

Raw ChIP-seq libraries from Stefflova *et al.*[558] were used for repeat content comparison to other two liver-specific transcription factors, CEBPA and FOXA1 for both mice subspecies. In each case, three biological replicates from different 8-week old male mice, and a matched control liver sample from another animal. Peaks found in at least two biological replicates out of the three were used for all downstream analysis. The raw FASTQ sequence were run through the same pipeline outlined earlier for peak calling. Interspecies comparison and repeat masking as described above for CTCF.

For studying the correlation between repeat content and the signal intensity of the TF binding site, all datasets for each transcription factor/CTCF in both subspecies were subsequently divided into ten 10% bins based on descending intensity of the ChIP-seq signal for each of the three evolutionary classifications: *musculus*-common, BL6-specific and CAST-specific. The repeat content for each bin of TF binding sites was then determined using the methodology detailed previously in 2.2.2.3.

2.2.2.5 Cross-Tissue Analysis of subspecies-Specific CTCF Binding

CTCF ChIP-seq data for BL6 adult (8 weeks) male mice were retrieved from the ENCODE Project data repository[451] for 12 tissues: lung, bone marrow, bone marrow macrophages, cortical plate, cerebellum, heart, kidney, thymus, spleen, olfactory bulb, small intestine and testis. We additionally used ENCODE libraries for the liver as a technical replicate to identify CTCF binding sites common in multiple tissues. In each case, two biological replicates from different 8-week old male mice, and a matched control tissue sample from another animal. Peaks found in both two biological replicates were used for downstream analysis. The raw FASTQ sequence data were used and run through the same pipeline outlined earlier for peak calling.

We used the overlap between our liver-derived CTCF peaks and the peaks from every ENCODE tissue to determine the tissue-sharedness of *musculus*-common/BL6-specific binding sites, using BEDTools intersect 2.2.5.0 with the options -wa -wb. UpSet plots were generated using the ComplexHeatmap package in R[571].

To calculate ENCODE tissues CTCF diversity index, we used the log10 of the p-value at peak summit computed by MACS version 2.1.0 during the peak calling step. For each liver-derived CTCF peak, if that peak was bound in an ENCODE tissue, the p-value for the binding of CTCF was retrieved. These values were subsequently used to calculate the Shannon Diversity Index for each tissue using Vegan package in R[572]. CTCF occupancy conservation across tissues was calculated as the fraction of CTCF peaks whose occupancy is conserved within each bin of Shannon diversity index.

Based on the results of the ENCODE tissue analysis, BL6-specific sites were then defined as tissue-shared or tissue-specific. Tissue-shared sites were CTCF binding sites found to be the intersection of all BL6-specific binding sites from the top four ranking tissue, plus the ENCODE liver technical replicate. All other BL6-specific sites were deemed tissue-specific.

2.2.2.6 Chromosome 4 Interferon-zeta gene-cluster Analysis

CTCF binding sites coordinates in *bed* format along with ChIP-seq coverage reads in those regions were uploaded for display on the Ensembl genome browser release 89[25]. These included reads from liver and the other four tissues, plus ChIP coverage reads from two histone marks for liver, H3K4ac27 and H3K4me3, from ENCODE data repository. Ensembl genome browser was used to display gene annotations, pairwise alignments with CAST and Rat, repeat elements enrichments for transposons and LTRs and genomic annotations. Sequence similarity for the 15-gene cluster, upstream CTCF binding regions, and the complete 15 constructs of CTCF binding sites plus the gene sequence plus $\pm$500 bp were determined using Clustal Omega[573], using default parameters.

We utilised the Comparative Genomics tool of the Ensembl Genome Browser to look at the BLASTz/LASTz whole genome alignment between the Chromosome 4 Interferon-zeta gene-cluster and all available pairwise alignments with other organisms[574]. An orthologous gene was found in the pig genome whose target sequence matched 14/15 from the mouse cluster with Query %id of $> 50\%$. We used BLAST/BLAT to scan the pig genome for paralogues to the gene based on sequence similarity. As with the mouse cluster, the Ensembl genome browser was used to display gene annotations, repeat elements enrichments for transposons and LTRs and GERP scores. Next, we scanned the 1 kb sequences upstream of each gene's TSS for the enrichment of motifs using MEME (4.12.0), setting 5 as a maximum number of motifs, and a motif width between 6-50 bp. The top 5 motifs from all upstream sequences were subsequently submitted to TOMTOM (2.14.0) to search available databases for annotated motifs to match[508, 509].

The manual annotation review of the locus on chromosome 4 determined that the annotation of the region was essentially correct with only a couple of minor issues identified and corrected. Specifically, Gm16686 was identified as a spurious protein-coding gene that will be removed in the next GENCODE release and has already been removed from RefSeq. RP23-400P11.4 was added as novel interferon pseudogene located at BL6 chr4:88754471-88754678 due to the clear pseudogenic characteristic of a significantly truncated 3' end. Finally, we reviewed Gm13286, which is annotated as a pseudogene in GENCODE, but considered protein-coding by RefSeq. It has a premature STOP compared to other family members, though it only loses the last 3aa of the typical protein. Based on the GENCODE annotation guidelines, Gm13286 is correctly annotated as a pseudogene although the coding status should be further investigated to make a definitive determination.

2.3 Results

2.3.1 CTCF binding is highly conserved, yet a considerable degree of occupancy is subspecies-specific

We performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) of CTCF in liver samples from BL6 and CAST (Methods). Comparable numbers of binding sites were obtained from all replicates in both subspecies. We limited our analyses to CTCF bound locations that yielded higher enrichment of signal when compared to control in at least two out of three of the biological replicates. Overlap analysis of binding sites based on the pairwise alignment of both subspecies revealed that most CTCF binding between CAST to BL6 was found at orthologous locations, with over 80% (>32,000) of the CTCF binding sites in common (*musculus*-common). This is consistent with previous reports across mammalian and rodent lineages[238, 575]. Nevertheless, we were able to determine that $\sim$ 20% of binding in either subspecies was found to be subspecies-specific, with approximately 7,000 of these subspecies-specific sites in BL6 compared to over 4,000 sites in CAST (Figure 2.1b). The disparity in the number of subspecies-specific sites between BL6 and CAST could be an artefact of the difference in CTCF peaks between them (Figure 2.1a).

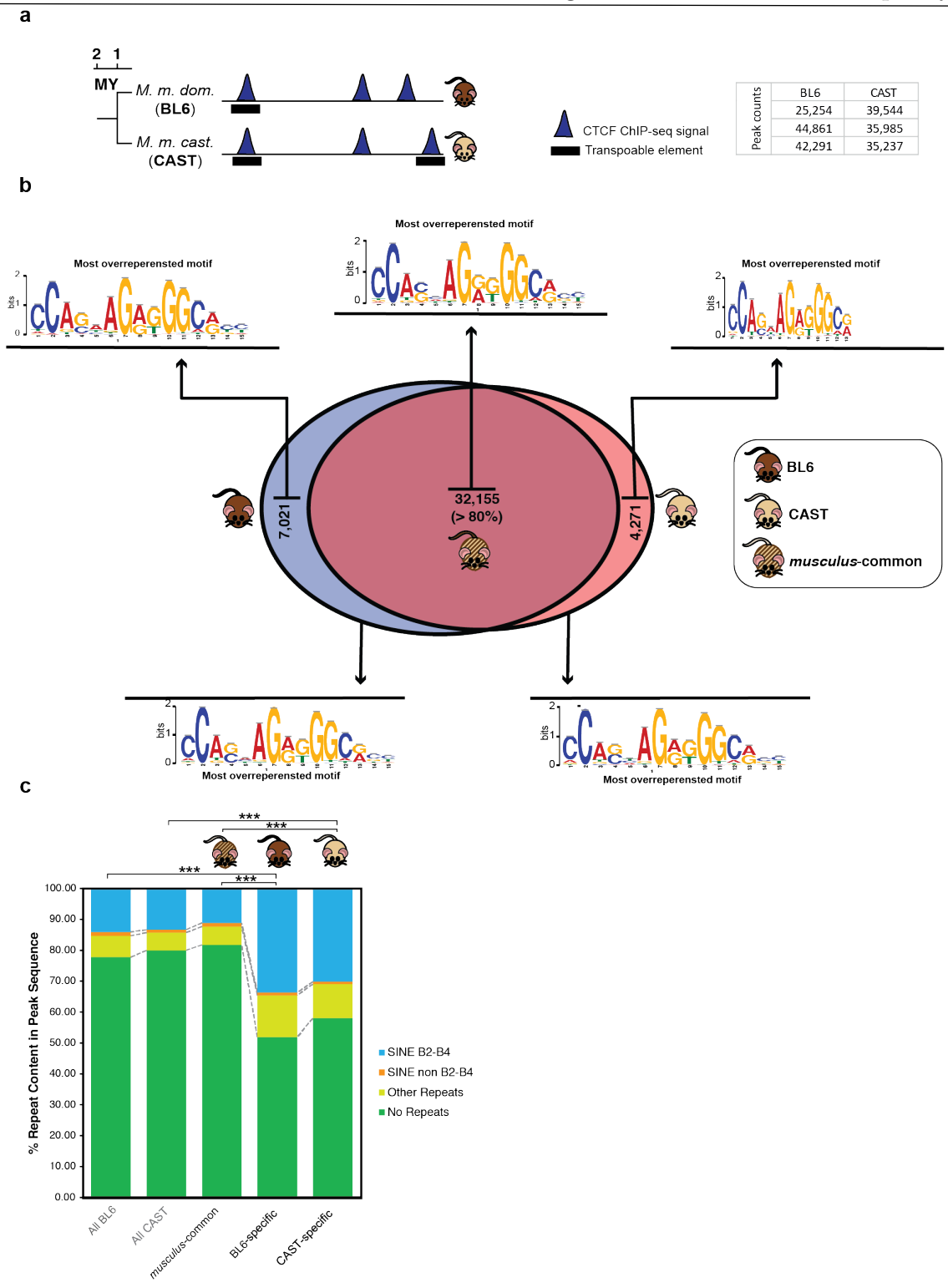


Figure 2.1: Identification of specific-specific CTCF binding in the genomes of BL6 and CAST mice.

a A schematic example of the contribution of transposable elements novel subspecies-specific CTCF binding. Approximate divergence times in millions of years are displayed on the left-hand side of the panel[576]. The peaks represent CTCF binding as determined from ChIP-seq experiments, while the boxes denote transposable elements. The table shows the peak counts (binding sites) retrieved from the three biological replicates for each subspecies. All downstream analysis utilised peaks common to a minimum of two replicates. **b** Overlap analysis of ChIP-seq peaks sub divides CTCF binding into five categories: Total BL6, BL6-specific, Total CAST, CAST-specific and *musculus*-common. The numbers inset are the binding sites in two subspecies-specific and the *musculus*- common sets. The most common motifs are indistinguishable among each set. **c** Bar plot showing the fraction of sequence annotated as repeats in each category described in **b**. The asterisks indicate the significance of enrichment of SINE B2-B4 elements between subspecies-specific sets and the entire set of binding sites for either subspecies and the *musculus*-common set (binomial tests with Bonferroni's correction, ***p-value < 0.0001).

Motif analysis was performed on the CTCF sites retrieved by peak-calling from each subspecies. In both subspecies, the canonical CTCF binding motif was obtained regardless of whether we considered all binding sites, or the subspecies-specific subset of sites, or those that were common to both subspecies (Figure 2.1b).

2.3.2 SINE repeat expansion is major driver of subspecies-specific binding

CTCF binding site evolution is known to be driven by repeat element expansion, particularly of the Short-Interspersed Elements (SINEs) superfamily of transposable elements (TEs)[269, 384]. We, therefore, investigated transposable element enrichment in the CTCF binding sites. The repeat content in all CTCF binding sites for either subspecies was comparable between the two subspecies, with an average of 21% of all sequences bound by CTCF embedded in repetitive elements (Figure 2.1c). Two-thirds of these sequences occurred within SINE TEs. With most of the binding sites shared between the two subspecies, it was unsurprising that the *musculus*-common binding sites have the same repeat composition (Overall: 18%, of which SINE: 12.4%) (Figure 2.1c).

However, the subsets differed substantially when the analysis was limited to the 20% of CTCF sites bound in a subspecies-specific manner. Compared to the *musculus*-common binding sites, the subspecies-specific sites were enriched in transposable and repetitive elements. Indeed 34% of BL6-specific CTCF sequences occurred within a SINE element, a more than a two-fold increase (BL6-specific vs. All BL6 binomial test

p-value=3.12766E-07, BL6-specific vs. *musculus*-common binomial test p-value=8.96592E-10). Comparable results were seen with the CAST-specific binding sites (CAST-specific vs. All CAST binomial test p-value=7.41176E-06, CAST-specific vs. *musculus*-common binomial test p-value=1.85083E-07) (Figure 2.1c). The enrichment of repeat content in subspecies-specific CTCF binding sites may be an underestimation of the true contribution as the approximately 15% of sites that were identified in only one replicate—and were thus left out of further analysis—also exhibit increased enrichment in repeat content.

When the *musculus*-common and BL6-specific sets of CTCF binding sites were compared to the four most common TE superfamilies in the mouse genome, CTCF was found to be highly enriched with SINE TEs at 71% for *musculus*-common sites, rising up to 76% for subspecies-specific sites for all sequences embedded in repetitive elements ($\chi 2$ test, p-value $<$ 2.2e-16). This constituted a three-fold increase over both randomised genomic regions matched for number and size, and the overall genomic background occupied by TEs in the BL6 mouse (Figure 2.2a). Conversely, longer repeat elements, from the LINE and LTR superfamilies, were clearly depleted for CTCF binding sites ($\chi 2$ test, p-value $<$ 2.2e-16), with subspecies-specific sites showing a slightly lower level of sequences occupied by LINE elements (7% vs. 9%). Whilst CTCF sites that are *musculus*-common showed a slight enrichment for DNA-transposons compared to randomised regions and the background (4% vs 2.8-3%), they were almost absent in subspecies-specific sites (~1%) (Figure 2.2a).

In addition to the observations in Figure 2.1c, CTCF binding site sequences were not only bound in SINE repeats more often than their *musculus*-common counterparts, but for each subspecies-specific binding site embedded in a SINE TE, most of the sequence is SINE-derived (Figure 2.2e). The majority of these BL6-specific sites were nearly all masked by SINE repeats (median = 77% of sequence), compared to either *musculus*-common site (39%) and randomised regions (31%). These differences in repeat element contributions to the sequence of CTCF binding sites were found to be highly statistically significant (Man Whitney U test, both p-value $<$ 2.2e-16).

We next quantified the age of all TE-derived sequences, estimated by the sequence substitutions of the repeat elements within binding sites, in *musculus*-common, subspecies-specific and randomised genomic regions. We used the percentage of mismatch in TE sequence from the consensus as a measure of the relative age of the TE element. Using sequence mismatches to estimate the relative age of repeat elements, evolutionary young sites inserted into new genomic positions via repeat expansion of TE elements have only had a relatively short evolutionary time to accumulate mutations. Conversely, binding sites in TE elements characterised by increased level of mismatches in the sequence originate from much older repetitive

sequence. Using this rationale, and in agreement with the results reported above, CTCF BL6-specific sites have a significantly lower levels of sequence mismatches in their TE-derived sequences (median = 17%) than either *musculus*-common sites (22%) or randomised genomic regions (21%) (Man Whitney U test, both p-values < 2.2e-16) (Figure 2.2b). This indicates the recent evolutionary origin of these sites in comparison with either conserved or randomised regions. Most of these evolutionarily young sequences originate from expansion of SINE repeat elements as evidenced by the restricting the analysis on the subset of CTCF binding sites within SINE TEs. These sites show a significantly younger age, illustrated by the low levels of sequence mismatch in their sequences (median = 16% for BL6-specific versus 22% and 24% for *musculus*-common and randomised genomic regions respectively) (Figure 2.2c). This recent cluster of SINE-derived CTCF sites is evidence of post-divergence expansion of the binding site that continued in each mouse lineage separately, and may yet still be ongoing. As with general TEs, these differences in SINE elements ages were statistically significant (Man Whitney U test, both p-value < 2.2e-16).

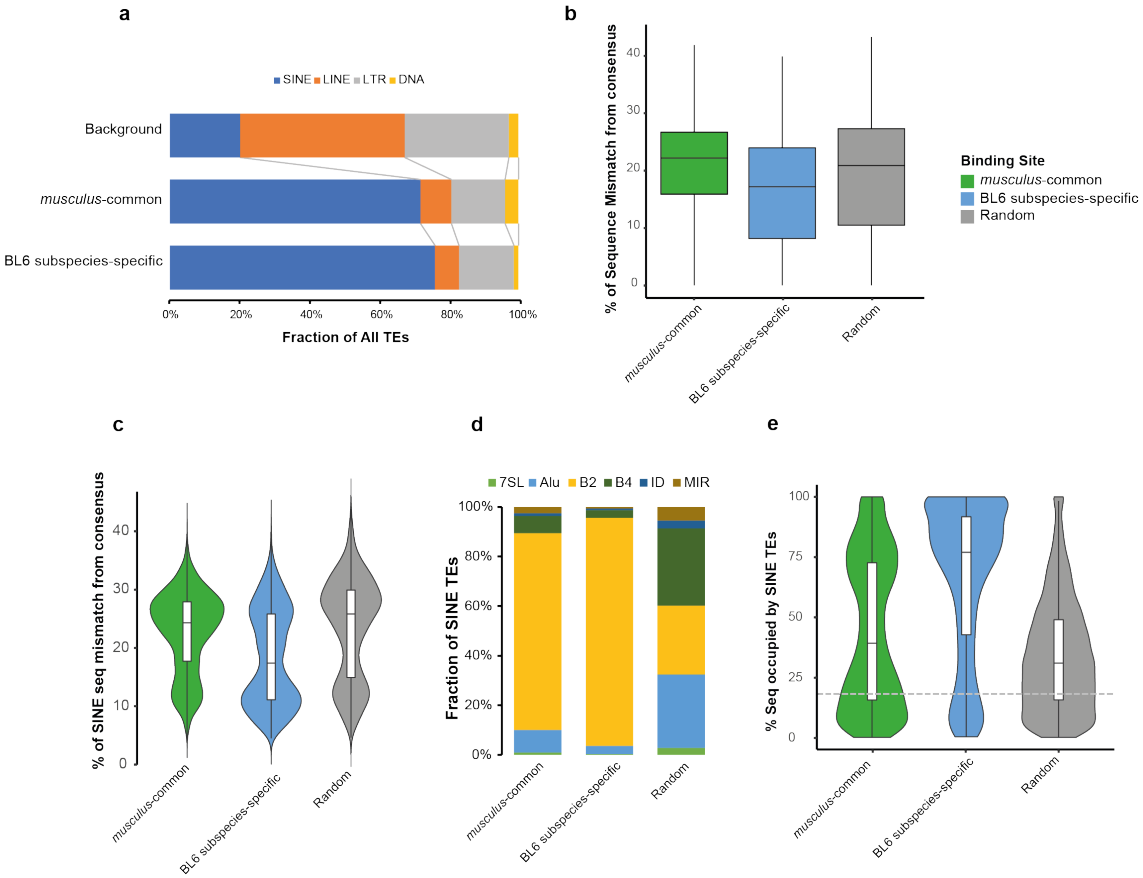


Figure 2.2: SINE transposable elements drive CTCF subspecies-specific binding

a Horizontal bar chart shows the proportion of major TE superfamilies in CTCF binding sites that are masked by repeat elements. The top bar represents the percentage each TE superfamily occupies in all repeat masked sequences in the BL6 mouse genome as a background. The bottom bar shows the fractions of these superfamilies in a matched set of randomised regions of the same length. **b** Box plot of the percentage of sequence mismatches/substitution from the TE consensus sequence of the all TE superfamilies in the different evolutionary categories of CTCF sites, compared to a matched random set. **c** Violin plot of the proportion of sequence mismatches/substitution from consensus in SINE TEs for the same categories from **b**. **d** 100% stacked bar plot of the proportions of the most common families of SINE TEs in all sequences masked by SINE elements in the different types of CTCF binding sites (The Alu family in the panel refers to the Alu/B1 rodent-specific family). **e** Violin plot of the proportion of sequence masked by SINE repeat elements in conserved/subspecies-specific CTCF sites, compared to the matched random set. The boxplots within each violin plot show the variation in extent of sequence masking for each category. The dashed grey line denotes the average genomic sequence occupied by SINE TEs in all TE-masked sequences of the mouse genome.

Most SINE-derived CTCF sequences in mouse belong to the B2-B4 rodent-specific family[269, 384, 409]. Analysis of SINE families make up in CTCF binding sites and randomised region showed that this is indeed the case, particularly with BL6-specific sites in which a further enrichment of the B2-B4 elements was evident (95% compared to 86% in *musculus*-common) (Figure 2.2d). However, our results provide evidence that the timing this CTCF binding expansion, previously known to be mouse- or rodent-specific, is in fact subspecies-specific and continued after the divergence between BL6 and the closely-related CAST subspecies. Even though SINE-derived *musculus*-common CTCF binding sites were made up of mostly B2-B4 elements, they appear to have been primarily involved in the evolution of subspecies-specific CTCF binding sites ($\chi 2$ test, p-value $<$ 2.2e-16).

2.3.3 CTCF binding sites have distinctive repeat profiles

For comparison, the evolutionary and repeat content analyses of *musculus*-common versus subspecies-specific binding sites was analogously performed for the liver-specific transcription factors CEBPA and FOXA1. First, ChIP-seq analyses of publicly available libraries for these two factors [558] were done to identify *musculus*-common and subspecies-specific binding in BL6 and CAST as described above for CTCF (see Methods). In both subspecies and for both transcription factors, repeat-based elements contributed more to specific-specific than to common binding, but to a lesser extent than that for CTCF. (Figure 2.3a *left*). However, unlike CTCF, few

SINE B2-B4 elements were bound by the tissue-specific transcription factors. Instead the repeat content in all sets of binding sites for both transcription factors originated from the various superfamilies of repeat elements with no obvious major contributor (Figure 2.3a *right* and 2.3b).

Taken together, comparison with the other two transcription factors showed that the expansion of subspecies-specific CTCF binding sites by SINE B2-B4 elements is not a general trend seen in liver-specific transcription factors.

We next asked whether there was a correlation between the repeat content of a CTCF binding site and its signal intensity as measured by the number of ChIP reads. From the subspecies-specific and *musculus*-common sets, binding sites were sorted by their signal intensity and then collected into 10 equal-sized bins. In the *musculus*-common set of CTCF binding sites, the repeat content increases as the intensity of the binding site decreases, In contrast, in the two subspecies-specific sets both the overall repeat content and the number and percentage of SINE B2-B4 elements remain roughly constant, regardless of the ChIP-seq signal intensity (Figure 2.3b *left*). Thus, the repeat content in both subspecies-specific sets is independent of the signal strength of the binding site.

Repeating this analysis for the two tissue-specific transcription factors revealed that the overall repeat content does not change with the ChIP-seq signal intensity, but remains roughly constant (Figure 2.3b *middle* and *right*). The contribution of SINE B2-B4 subfamily to this is negligible (~2-3%) (See Appendix 1).

The relationship between ChIP signal intensity and repeat content supports the expectation that CTCF subspecies-specific binding sites have recently arisen from SINE B2-B4 elements and thus have a different evolutionary origin than binding sites of typical tissue-specific transcription factors. However, the sites with low signal intensity and observed in multiple replicates are not mere sequencing noise, because the CTCF motif was retrieved from these sites, and their genomic distribution is indistinguishable from those CTCF binding sites with higher intensity (Figure 2.1b). Furthermore, though not SINE-driven, we showed that liver-specific TF binding sites are also actively evolving in those two subspecies since the time of their divergence. In sum, these results illustrate the power with which transposable elements rapidly modulate transcription factor binding, shaping nearly 50% of the subspecies-specific CTCF occupancy profile in just one million years of divergence time between BL6 and CAST.

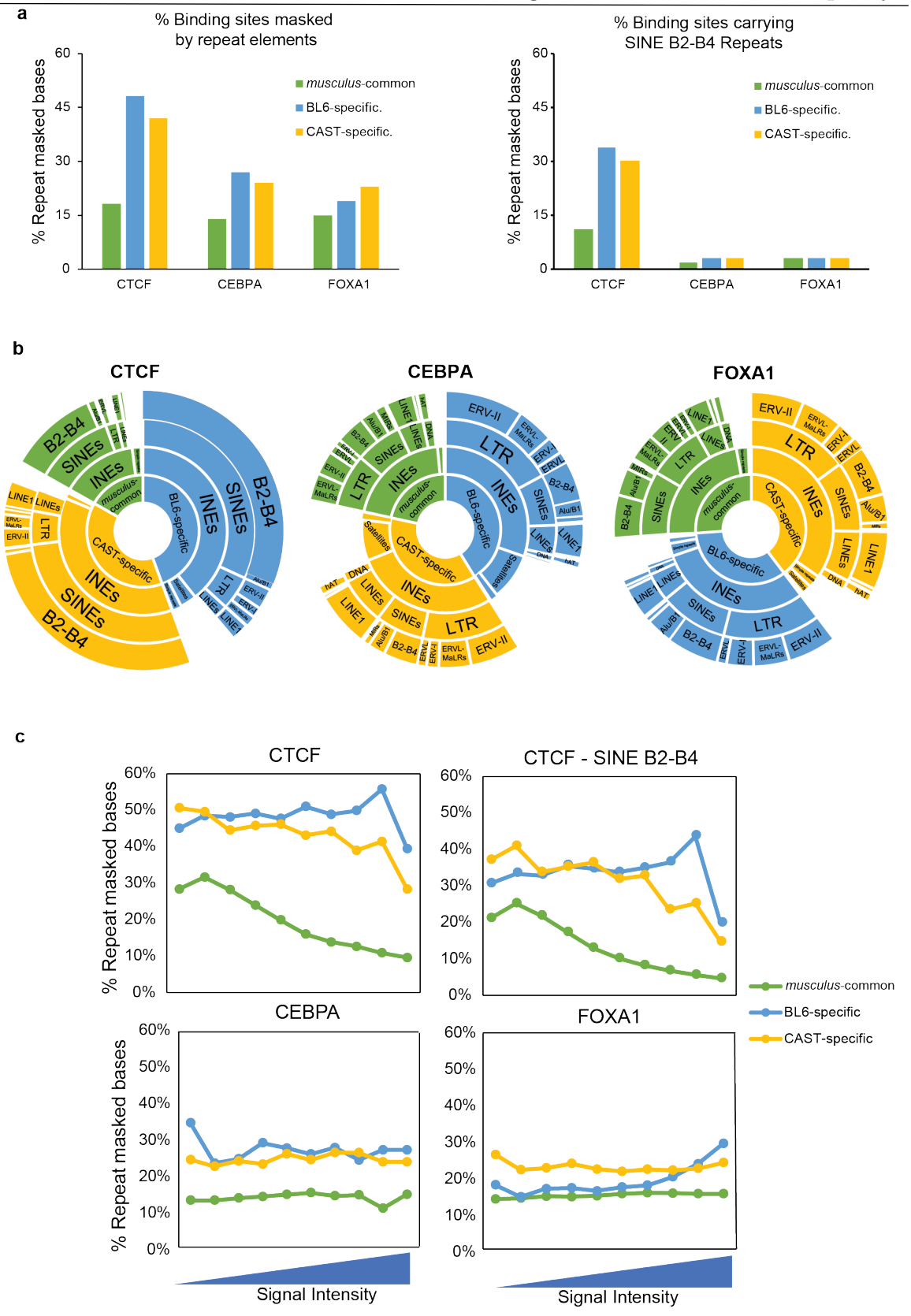

a
% Binding sites masked by repeat elements
% Repeat masked bases
musculus-common
BL6-specific.
CAST-specific.
CTCF
CEBPA
FOXA1
% Binding sites carrying SINE B2-B4 Repeats
b
CTCF
CEBPA
FOXA1
c
CTCF
CTCF - SINE B2-B4
CEBPA
FOXA1
% Repeat masked bases
musculus-common
BL6-specific
CAST-specific
Signal Intensity

Figure 2.3: CTCF binding sites have distinctive repeat profiles as compared to tissue specific transcription factors

a For CTCF, CEBPA and FOXA1, subspecies-specific binding sites are enriched in repeat elements (left) (binomial test, all p-values $<$ 2.2e-16), but only CTCF is associated with the SINE B2-B4 sub-family (right). **b** The characterisation of the different categories of repeat elements in the binding sites of CTCF and the two TFs from (**a**) shows that the CTCF has a distinctive TE profile. INEs stands for Interspersed Elements. Whilst the types of TE in the binding sites of CEBPA and FOXA1 widely vary between them, and within their binding sites depending on their evolutionary status, SINE B2-B4 elements almost exclusively make up all TE-derived occupancy in CTCF regardless of subspecies-specificity. The sizes in each plot are proportional to the sequence occupied by each type of repeat element, and their overall proportion of the total binding site sequence masked by TEs according to their evolutionary status. **c** Subspecies-specific CTCF sites contain a consistent level of repeat annotated sequence regardless of ChIP-seq signal intensity. More intense *musculus*-common binding sites have a lower fraction of repeat annotated sequence. The same pattern is observed for the subset of CTCF sites within SINE B2-B4 elements. The x-axis is arranged in 10% bins based on binding site signal.

2.3.4 A subset of BL6-specific CTCF binding is tissue-shared.

Although CTCF is known to be bound across multiple-tissues, the numbers and locations of its binding sites can vary among tissues[577]. We, therefore, investigated the association between subspecies-specificity and tissue-specificity by profiling the pattern of CTCF binding in tissues other than the liver using publicly available data. We used ENCODE CTCF ChIP-seq data for BL6 adult (8 weeks) male mice from 13 tissues: liver, lung, bone marrow, bone marrow macrophages, cortical plate, cerebellum, heart, kidney, thymus, spleen, olfactory bulb, small intestine and testis[451]. Since ENCODE did not perform ChIP-seq analysis on CAST samples, all tissue analyses were limited to BL6 only.

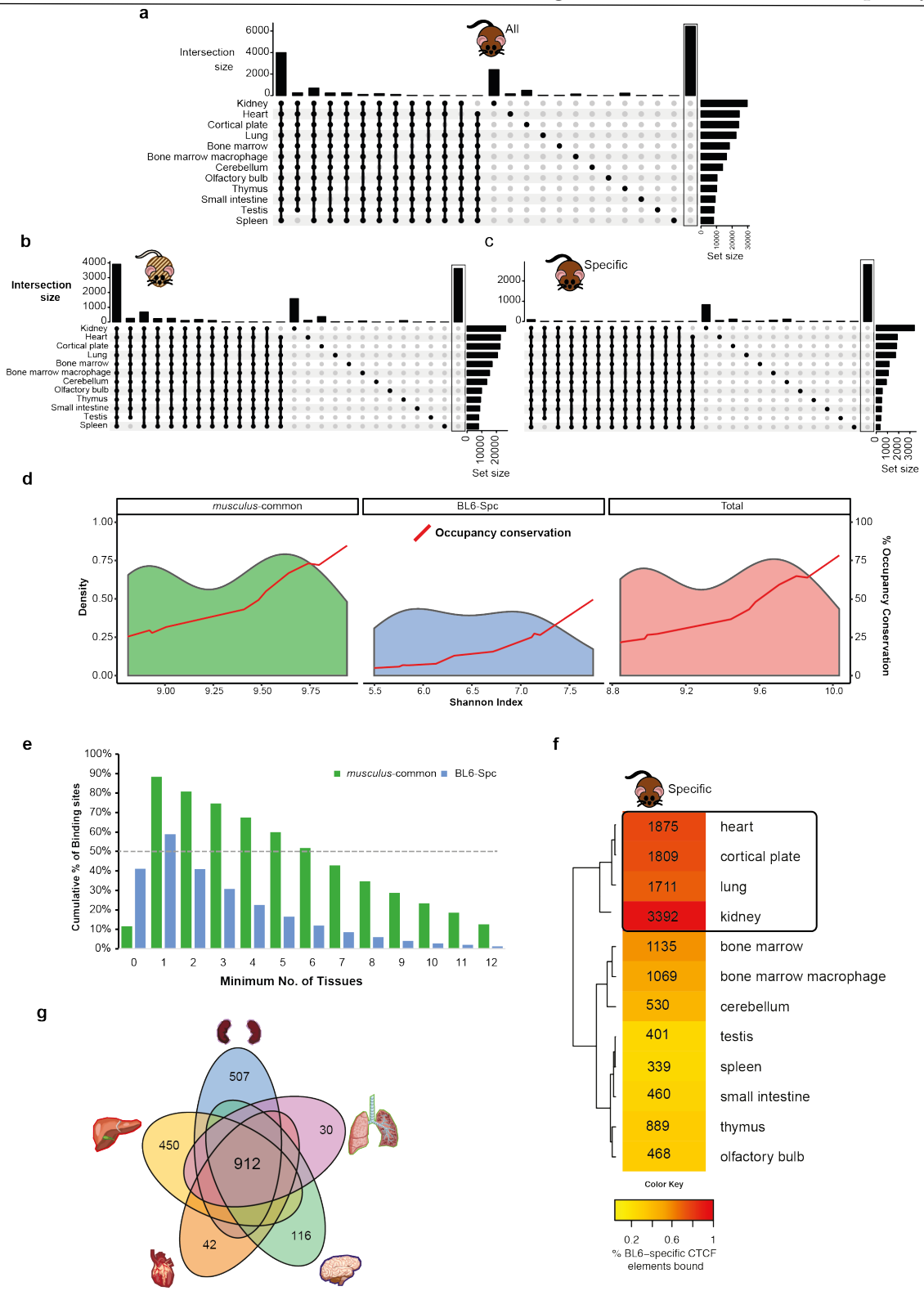


Figure 2.4: Almost a 1000 BL6 subspecies-specific CTCF binding sites are shared among five tissues.

a, b and c UpSet plot of the liver-derived CTCF binding sites found across the 12 mouse ENCODE tissues for all sites (**a**), conserved (**b**) and BL6-specific (**c**). The number of sites bound at each combination of tissues is indicated on the y-axis on the top bar chart. The original plot was reduced to these 26 combinations representing only highly tissue-shared and tissue-specific. The rightmost bar on each UpSet plot (*boxed*) indicates the number of CTCF binding sites that were not found to be bound in any other ENCODE tissue library. **d** Density plots of the association between CTCF binding strength in *musculus*-common/BL6-specific sites and occupancy conservation across 12 tissues. The plots display the frequency of CTCF shared binding across tissues. Diversity values are indicated on the x axis as calculated using Shannon Diversity Index from the p-value estimates of peak calls of each category (see Methods). The red line is for the proportion of conserved CTCF occupancy within each bin of Shannon index, calculated based on the number of CTCF sites bound for each category across tissues separately. **e** Bar plot of the proportion of CTCF binding sites bound in ascending number of tissues in conserved versus BL6-specific sites. The y axis represents the cumulative percentage of binding sites found at the minimum number of tissues on the x axis. The dashed grey line denotes the minimum number of tissues at which 50% sites are shared. **f** Overlap of CTCF binding from 13 ENCODE Project derived data sets with our liver-specific BL6 data. The four tissues with the highest overlap are enclosed and used for further analysis. The number of peaks shared with each tissue are inset. **g** Number of BL6-specific CTCF binding sites shared among subsets of the four selected tissues, plus the ENCODE liver as an added technical replicate.

Analysis of ENCODE tissue libraries of CTCF showed that at least 10.5% of all binding sites, are bound in all ENCODE tissues (almost 4000 sites), and over 1800 more have their occupancy conserved in a minimum of 11 tissues (Figure 2.4a). On the other hand, 17% of all CTCF sites appear to be liver-specific, with no shared occupancy in any other ENCODE tissues. Of all ENCODE tissues analysed, kidney appears to have the highest degree of tissue-shared binding with the liver, with more than 78%, of which over 2000 sites are bound exclusively between the two tissues (Figure 2.4a). When we stratified these sites based on their evolutionary origin, the patterns above were mirrored in the *musculus*-common set of CTCF binding sites; 98% ($>$ 3900) of all CTCF sites bound in all 12 tissues were *musculus*-common (Figure 2.4b). The results from the BL6-specific set of CTCF sites were, on the other hand, to the contrary. Slightly over 1% of all BL6-specific CTCF sites were bound in all 12 tissues, and 41% of these sites ($>$2800) were found only in the liver (Figure 2.4c). The kidney, again, appeared to be the tissue with which most occupancy is shared, albeit greatly reduced now from 85% in *musculus*-common to just about 50% in BL6-specific sites (Figure 2.4c).

In addition to looking at shared CTCF occupancy in other tissues, we evaluated the strength of CTCF binding sites in all of the 12 mouse ENCODE tissues, using the Shannon Index[578] as a measure of their diversity of evolutionarily variable CTCF binding on the basis of their abundance and conservation. The expectation is that sites that are tissue-shared are more likely involved in regulatory functions, hence they are under increased selective pressure that keeps their levels of shared binding high tissue-wide[399].

Results confirmed the findings above, showing high Shannon index values across tissues, correlated with a great degree in CTCF occupancy conservation (Figure 2.4d *rightmost panel*). These results were more strongly observed in *musculus*-common sites, with higher density at higher values of the diversity index (Figure 2.4d *leftmost panel*). The Shannon index high values distribute smoothly in a bi-modal trend, with the bottom five tissues from Figure 2.4a and b clustering together towards the lower range of the diversity index, and the top 5 tissues occupying the cluster at the higher end of the curve. The diversity curve for the BL6-specific CTCF sites was, however, markedly different. The calculated tissue index values were much lower and extended in a wider scale, flattening the distribution, a further sign that subspecies-specific CTCF binding is predominantly tissue-specific, its occupancy in other tissues is far more restricted.

In light of the analyses performed above, we explored the possibility of finding a subset of BL6-specific CTCF binding sites with elevated levels of tissue-sharedness. We theorised that increased tissue-permeation to CTCF subspecies-specific binding could be a precursor to their adopting functional roles. We first looked at how many of these sequences are found in progressively more tissues (Figure 2.4e). Analysis of ENCODE tissue data showed that whilst a minimum of 50% of all *musculus*-common CTCF sites are found in at least 6 tissues, the same proportion of sites can be found in only one other tissue for BL6-specific sites. The analysis; however, suggested that 16% of subspecies-specific sites can be tissue-shared in a minimum of 5 tissues (Figure 2.4e).

We identified the top five ENCODE tissues by the number of CTCF binding sites that co-occur with our liver BL6 ChIP-seq datasets for further analysis. As expected, ENCODE liver and kidney have the most overlap with our liver datasets (Figure 2.4f). For the BL6 binding sites we identified as *musculus*-common, 67-85% are shared in these five tissues and 26-49% of the BL6-specific sites we identified in liver are also bound in these five ENCODE tissues. The analysis only used CTCF binding sites that were retrieved from at least two ENCODE biological replicates, making the number of estimated binding sites, especially those that are BL6-specific, likely conservative (Methods).

Focusing only on the five ENCODE tissues most similar to liver in CTCF binding profile (Figure 2.4f *enclosed*), we were able to identify a subset of CTCF subspecies-specific sites that are bound in all five tissues. There were 912 CTCF sites found in our data and shared with ENCODE liver, kidney, heart, lung and cortical plate (Figure 2.4g), spread over all mouse chromosomes. These sites constitute 13% of all of our BL6-specific CTCF sites, but we hypothesise that shared binding in these different tissues may indicate an increased involvement in genomic functions compared to their tissue-variable counterparts.

2.3.5 Tandem duplication event of BL6-specific CTCF binding sites linked to the expansion of an interferon gene cluster.

While investigating the subset of BL6-specific and tissue-shared CTCF sites genomic distribution, we uncovered a single TAD on chromosome 4, band C4, that contains a cluster of 15 BL6-specific, tissue-shared CTCF binding sites within a 58 kb window. A CTCF binding site precedes the TSS of each of 15 copies in a cluster of type 1 interferon zeta (Ifnz) genes. No *musculus*-common CTCF binding sites were found inside this cluster of genes, whether in genic or intergenic regions, and all nearest *musculus*-common sites are scattered in no particular pattern or clustering (Figure 2.5 *Top*). All 15 CTCF binding sites collocated with cohesin, were of similar lengths and exhibited comparable read coverage signal in every tissue. This genomic cluster is contained on a single clone within the reference BL6 mouse genome assembly (https://www.ebi.ac.uk/ena/data/view/AL928605) and is thus unlikely to be the result of a genome assembly artefact.

The 15 genes in this uncharacterised cluster are all from the type 1 interferon family of genes, for which many other members are present in and around the same locus. Although many of the upstream and downstream interferon genes have been previously described[579, 580], this 15-gene cluster, though previously reported[581], is yet to be functionally characterised. All but one have putative gene names, and two are annotated as pseudogenes. We have manually reviewed the annotation for all of the genes in the cluster and found it to be sound (see Methods). Specifically, the majority of the genes are novel or predicted protein coding genes in the GENCODE annotation[582], and all have evidence on the transcript level[583, 584]. The Gene Ontology[585, 586] terms most associated with these genes are for cytokine activity and type 1 interferon receptor binding molecular functions, with involvement in adaptive immune response in the defence mechanisms against viral infection.

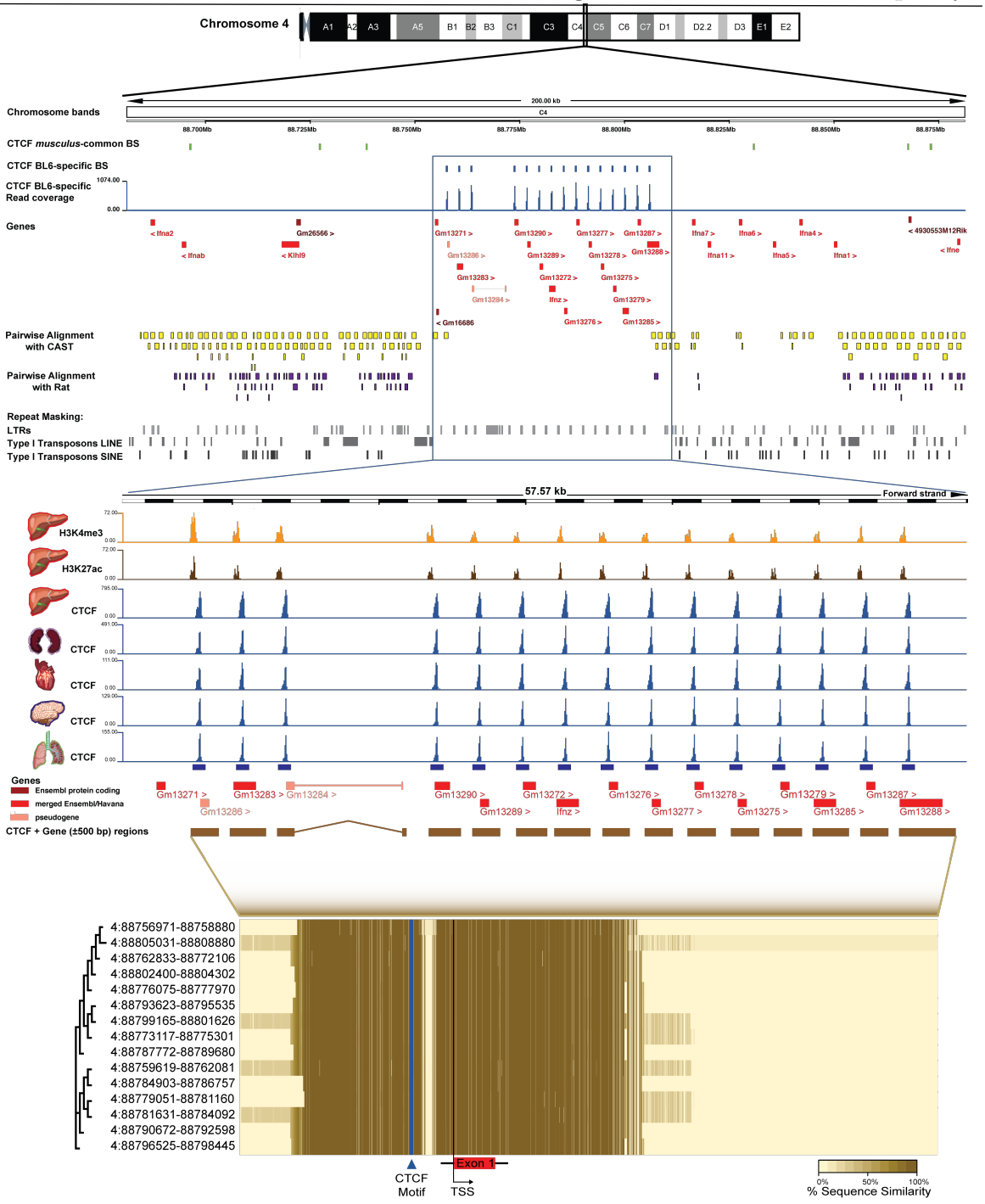


Figure 2.5: Evidence of a tandem duplication event of BL6-specific CTCF binding sites on Chromosome 4 in multiple tissues linked to the expansion of a family of interferon genes.

Top: A zoomed out summary view of 200 kb of Chromosome 4 band C4. The tops two tracks show the CTCF-Cohesin regions in *musculus*-common and BL6 tissue-shared sites. The next track in blue indicates read coverage signal from

CTCF BL6 tissue-shared binding sites. Below that, a summary of all genes in the 200 kb window. The arrowheads next to the gene names indicate the direction of transcription. The next two tracks in yellow and purple are pairwise alignment of the region between BL6 and CAST and Rat respectively, showing the noticeable lack of any orthologous regions in either subspecies. The bottom three tracks in grey shading illustrate the repeat content of the whole genomic region with the noticeable lack of any large scale repeat elements in the highlighted region.

Middle: A zoom-in view of the 57.6 kb region 4:88752534-88810107 in which CTCF-Cohesin BL6-specific, tissue-shared binding was observed. The top two tracks in orange and brown indicate read coverage signal from H3K4me3 and H3K27ac respectively. The next five tracks in blue indicate CTCF ChIP-seq read coverage in each of the five tissues discussed in Figure 2.4. The corresponding 15 genes are shown below the tracks.

Bottom: A heatmap of sequence similarity in the multiple sequence alignment of the 15 CTCF-Cohesin binding sites on the C4 band of BL6 chromosome 4. The numbers denote the start and end positions of each binding site. The heatmap also shows sequence similarity in the multiple sequence alignment of the 15 genes on the C4 band of BL6 chromosome 4 that are all preceded by a CTCF-Cohesin binding site. The numbers denote the transcription start site and the name of each gene. The dendrograms on the left of each heatmap are clustering trees showing the relationship between the sequences based on their sequence similarity.

Although we identified this region by the presence of BL6-specific CTCF binding, the entire region containing the gene cluster does not, in fact, have an orthologous region in the CAST genome (Figure 2.5 *Pairwise Alignment*). Moreover, there is neither an orthologous region in Rat nor any of the other 13 mouse strains/species available in the pairwise alignment of mouse strains available in Ensembl release 91[25].

Strikingly, this cluster was characterised by the absence of transposon-driven repeat elements, with a complete lack of type 1 SINE or LINE transposons, despite them occurring both up- and down-stream of this region. Indeed, the only repeat elements within the 58 kb window were LTRs, all from the LTR-ERVK subfamily, and all either 450 or 550 bp in length (Figure 2.5 *Grey Tracks*). The LTR elements within the cluster generally occurred in intergenic regions with none in the CTCF bound regions or 500 bp up or downstream from the gene bodies, with the exception of one cluster of repeats in the intron of a pseudogene (Gm13284). Upstream CTCF binding sites were completely devoid of repeat elements with only six, short simple repeats in those 15 genes (ave. simple repeat length 50 bp).

Figure 2.6: Convergent evolution of an orthologous interferon gene cluster in pig.

a Genome browser display of BL6-Pig LASTz pairwise alignment of the 15-gene cluster. Pink tracks show the BL6 genome regions aligning to sequences in the pig genome. **b** A zoom-in view of the orthologous gene cluster of interferon precursors in the pig genome. The orthologous gene in (**a**) is shown as the leftmost gene in the window in brown italics. The 12 paralogues to this gene are highlighted in brown italics with the other interferon genes in light grey. The arrowheads indicate the direction of transcription. The dark grey tracks at the bottom indicate the LTR repeat content of the region. **c** A schematic diagram showing the position of the CTCF motif enriched at around 1 kb from the TSS

of 12/13 genes in the cluster, with the motif composition below the orange track indicating the position. The motif underneath is the CTCF canonical motif with the p-value of the probability that the match occurred by random chance.

Both the upstream CTCF regions and the genes in this cluster exhibit high sequence similarity, with near identical sequences for large portions of their lengths (Figure 2.5 *Heatmaps*). The sequence similarity also extends beyond the genic regions into the LTR repeat elements punctuating the intergenic distances between them. These characteristics suggest a tandem duplication event that repeatedly carried the ancestral gene, and copied its sequence along with upstream regulatory regions to produce the regulatory landscape of this cluster. We do not, however, find any evidence of SINE or LINE transposon activity in the duplication event. As these genes have a high degree of similarity with each other, and other interferon genes outside the cluster on chromosome 4, we could not discern the evolutionary history of this BL6-specific gene cluster. To assess whether these regions are regulatorily active, we used previously published histone modification data for H3K27ac and H3K4me3[451]. We observed that the CTCF binding sites co-located with both modifications (Figure 2.5 *Middle*), which taken together generally signify an active promoter signal[209].

Nevertheless, genomic region comparison with other eutherian mammals revealed that 14 of the 15 genes in the cluster align to a single gene in the pig (*Sus scrofa*)[587]. LASTz pairwise whole genome alignments show that all but one of the BL6 genes align with between 50-100% coverage to the ENSSSCG00000039987 gene, a novel predicted protein-coding gene on pig chromosome 1 (Figure 2.6a). There are 13 annotated paralogues to this gene in the pig genome that all lie within a 500 kb cluster, albeit separated by 24 intervening genes between them. All of these genes belong to the same Ensembl protein family in pig, PTHR11691 (Interferon Precursor), and all of the members of this family, except one, are within this cluster. Unlike the cluster found in BL6, however, the pig genes are evenly divided between reverse and forward orientation (Figure 2.6b). The pig cluster is also enriched for LTRs and these also punctuate the intergenic distance between the genes. The cluster has very low GERP conservation scores compared with no constrained elements, unlike the regions up- and downstream of the cluster. We used the 1kb flank of all 14 pig genes orthologous to BL6 genes for motif discovery and found the CTCF motif in 13 of the 14 genes (Figure 2.6c). Taken together, the observation of two clusters comprising related immune genes that are closely associated with subspecies-specific CTCF sites suggests a role for CTCF in immune response reflected here potentially via a convergent evolutionary process.

2.4 Discussion

While TE-derived expansion of CTCF binding sites is a well-documented model of the evolution of TF occupancy in several mammalian lineages[269, 384, 559, 588], our results newly reveal how rapidly expanded repeats can acquire distinct functional signatures, even between two closely related mammalian species.

In this chapter, we used two closely-related mouse subspecies, BL6 and CAST, separated by one million years of evolution to study the evolution of CTCF binding and the functional signature of evolutionary young sites. Our results demonstrate that many subspecies-specific CTCF binding sites are bound across multiple tissues that originate from different germ layers. This similarity of genomic behaviour hints at a potential functional role that they may play, either independently or as potentially redundant binding in case of the loss of a nearby *musculus*-common CTCF binding site due to mutation. Whether evolutionary young binding sites are capable of taking over the functions of ancestral sites or may be more associated with lineage specific function will require further investigation. The latter model finds support in a recent computational approach that showed 15% of lineage-specific transcription factor binding sites are enriched for genes involved in cell-type specific pathways, such as the fast-evolving olfactory pathway, and have distinct biological implications when compared to ancestral ones[589].

Our data provide insight into how frequently functional CTCF binding sites are conserved between subspecies and/or shared among tissues. Previous observations have shown that tissue-shared CTCF binding sites are more conserved than cell type-specific CTCF-binding sites[590] and, reciprocally, that cell-type specific CTCF binding sites are more likely to be lineage- or subspecies-specific than tissue-shared sites[589]. To illustrate the functional differences between evolutionary young CTCF sites, we determined the level of CTCF occupancy conservation across tissues. Our findings demonstrated that whilst the majority of these young sites exhibit significant tissue-restriction in terms of their occupancy when compared to the *musculus*-common set, a subset of sites do show consistent binding across several tissues belonging to all three germ-layers. Other studies have reported that TE-derived subspecies-specific binding in primates, tissue-specific in particular[446, 591], demonstrates the potential for regulation of gene expression. Taken together, our results define a hierarchy for CTCF sites: those that are both evolutionarily conserved and tissue shared are most likely to be functional, followed by those that are either conserved or shared, and then those that are subspecies and tissue specific.

We discovered that a cluster of 15 type 1 interferon zeta family genes, associated with retrotransposable expansion of LTR elements, where BL6-specific, tissue-shared

CTCF binding was observed just upstream of every gene in the cluster. The CTCF binding, co-localised with histone modifications indicative of active promoters, is an example of a region where subspecies-specific binding may have led to the introduction of novel transcriptional regulation. The LTR elements punctuating the intergenic distance between the genes of this cluster may have facilitated the observed tandem duplication via non-allelic homologous recombination, or helped derive novel binding site locations for transcription factors and regulatory elements[592].

Even though this gene cluster has previously been shown to derive from a lineage-specific expansion of IFNα-like gene in rodents only[581, 593], this gene cluster has not been well-described in the literature, and no functional annotation work has been done on any of its members. Flanking genes in the interferon locus have been characterised, however, including the identification of orthologues in the interferon locus on human chromosome 9[580]. Another study showed that some of the flanking genes had constitutive transcriptional activity at low levels in the absence of viral infection[579]. This latter study reported the tandem array of 16 consecutive genes we have characterised here, which they incorrectly thought to be an assembly artefact[579]. Another study looking at the evolution of IFNα reported a similar expansion of genes from the family between the BL6 and 129/5v mouse strains[579]. A potentially similar BL6-specific expansion has been observed in the Abp gene cluster although not involving CTCF and progressively correlated with LINE and LTR enrichment[455, 559].

The phenomenon may yet transpire to be more common than observed in mice. The IFNδ gene cluster found in the pig genome carried several similarities to the cluster in BL6. Although we do not have CTCF binding data in pig, the presence of the CTCF canonical motif less than 1000 bp of the TSS in almost all of the genes is indicative of possible involvement. Xu et al. [581, 593] reported that ancestral IFNα-like genes duplicated during mammalian evolution and segregated into subtype. Of these IFNα-derived interferon subtypes, IFNδ and IFNζ underwent convergent evolution, forming an outgroup as a result of similar selection pressures in different subspecies, namely rodents and pigs. Whilst IFNδ propagated in the porcine genome, while it failed to gain a reproductive response in the mouse, and vice versa for IFNζ[581, 593].

Our results demonstrate a set of evolutionarily young CTCF sites that have been captured into operational regions of the genome and apparently adopted similar functions to *musculus*-common CTCF sites. Previous work has shown that mouse and rat subspecies-specific CTCF sites are comparable to mammalian conserved CTCF site in their ability to demarcate chromatin domains and modulate transcription[269], but our results go further by focusing on a cohort of several thousand of the evolutionarily youngest CTCF binding sites that have arisen in just the last 0.5 MY. Indeed, the

subset of these sites shared across multiple BL6 tissues show the strongest and most convincing functional signatures.

Further investigation will be required to establish the phenotypic consequences of evolutionarily young and subspecies-specific CTCF binding sites. The exact biological functions of many of these subspecies-specific CTCF sites will have to be established through targeted deletions of particular sites *in vivo* or using conditional knockdown of CTCF in suitable cell-line as CTCF knockout is embryonic lethal. However, it is clear that even the youngest CTCF sites carry multiple functional signatures that are indicative of contribution to transcriptional regulation in one or many tissues. Other transcription factors may have evolved subspecies-specific regulatory functions in a lineage-specific manner either through duplication events or in consequence to transposable element activity. The expansion of subspecies-specific regulatory elements concurrent with lineage-specific gene clusters, as we show in this study, is likely to be a common feature in mammalian genome evolution.

Chapter 3

Pervasive effects of *trans*-acting variation on CTCF occupancy

3.1 Introduction

Regulatory variants that change gene expression levels can be broadly classified into two main categories based on their physical genomic location relative to the genes they regulate: (1) *cis*-acting regulatory variants which mediate differential expression in a direct manner by influencing the local genomic sequence such as variants in promoters, enhancers and *cis*-regulatory modules; (2) *trans*-acting regulatory variants which mediate differential expression through diffusible elements such as proteins (TFs) or ribonucleic acids (ncRNAs, eRNAs, etc.). These variants result in the divergence of gene expression levels between species (in *cis*) or within individuals of the same species (in *trans*)[260]. These two classes reflect the differences in the gene expression levels that their inheritance mechanisms result in, and the type of selective pressure they are subject to[395, 542, 594]. *Cis*- and *trans*-regulatory factors undergo distinct evolutionary trajectories, displaying various extents of pleiotropic effects, as has been experimentally reported for loci affecting gene expression[595].

The contribution of *cis* and *trans* regulatory variants on the divergence in gene expression has been investigated in two main methods at the genome-wide level. The first method uses expression quantitative trait loci (eQTL) to identify regulatory mutations. In this method, the total gene expression level is measured across a population, and genetic variants (single nucleotide variants, or SNVs) are genotyped for the same individuals, then correlated with the expression levels[517]. The other method to detect *cis* and *trans* variants is to compare expression differences between

species with expression differences between alleles in inter-species F1 hybrid progeny[544, 596]. This method has been recently used to characterise *cis* and *trans* regulatory potential in flies[542, 543, 596-599], yeast[539-541] and mice[257, 260, 600]. These studies have improved our understanding of regulatory divergence within and between species.

Direct mutations affecting the TF binding site are a rare occurrence[558, 601, 602]. Recent studies propose this to be explained by long-range TF correspondence or *cis*-acting variants near, yet outside, the core motif[603, 604]. Many *cis* variants are not the primary targets of natural selection as suggested by the observation that expression levels of genes are fine-tuned by *cis* variants, following regulatory changes in *trans* [260]. Wong *et al.* demonstrated that *cis*-acting variants are the main driver of TF occupancy divergence using a similar approach in three tissue-specific TFs: CEBPA, HNF4A and FOXA1[257].

CTCF binding is, on the other hand, strikingly conserved across hundreds of millions of years of evolution, suggesting that CTCF binding sites are under similar selective pressures as the coding sequences of genes, a feature unique among TFs[267]. However, genetic heterogeneity and cell type specificity drive inter- and intra-individual variation in the expression of CTCF in a variety of tissues[605, 606]. A study using human lymphoblastoid cells found that 7% of DHS sites and 11% of CTCF binding sites exhibit allele-specific effects, and another showed allelic bias in CTCF binding sites in analysis of footprints with predicted binding factors[607, 608]. A previous study using eQTLs to measure binding of CTCF in 51 HapMap cell lines identified 1000s of QTLs where genotype differences were associated with differences in CTCF binding intensity. Hundreds of these were subsequently confirmed by observable allele-specific binding bias. However, the majority of these loci were at least 1 kb from the CTCF binding motif[609].

To our knowledge, an investigation into allelic-specific CTCF binding in response to *cis*- and *trans*-acting variants using an F1 hybrid system has not yet been conducted. Here we leverage the methodology developed in Goncalves *et al.*[260] and Wong *et al.*[257] to dissect CTCF occupancy divergence in mammals using F1 hybrid mice from two closely related subspecies, separated by half a million years of evolution. We highlight that CTCF binding, unlike tissue-specific TFs, is influenced by *cis* and *trans* factors that mediate its binding in allele specific contexts. CTCF binding does not display measurable coordination of regulatory mechanisms with proximal or distal CTCF sites. CTCF lineage-specific *cis*/*trans*-influenced binding is not common. Furthermore, *cis*- and *trans*-acting effects on CTCF occupancy, though mainly additive, display detectable dominant effects. Taken together, these results elucidate the complex pattern of effects a tissue-wide TF like CTCF displays in response to

genetic differences in its binding site that is considerably different from tissue-specific counterparts.

This investigation is the result of a collaboration between Dr. Paul Flicek's research group at EMBL's European Bioinformatics Institute and Dr. Duncan Odom's laboratory at the Cancer Research UK Cambridge Institute. Dr. Bianca M. Schmitt performed most of the wet lab experiments for this project, Dr. Emily S. Wong ran the initial alignments and provided the (**`cistrans cat assignment.R`**) code, and I carried out the remainder of the computational analysis, except where otherwise specified.

3.2 Methods

3.2.1 Experimental methods

3.2.1.1 Animal breeding and sample collection

The experiments were conducted using mice from two subspecies: C57BL/6J (stock number: 000664, source: Charles River Labs) and CAST/EiJ (stock number: 000928, source: The Jackson Laboratory). These two mouse subspecies were used as parental F0, and were mated to breed for the reciprocal crosses of the F1 mice. All biological replicates collected for the purposes of this investigation were sampled from adult male mice, 8-12 weeks of age, and harvested between 8 and 11 a.m. All animals were kept in similar husbandry conditions in the Biological Resources Unit of the Cancer Research UK–Cambridge Institute under a Home Office Licence.

Sampling of liver by perfusion was done on mice post-mortem, followed by tissue dissection. Harvested tissue samples were quickly chopped and transferred into a cross-linking solution with 1% formaldehyde in preparation for ChIP-seq protocol. Tissue samples were incubated for 20 minutes before quenching with 1/20th volume of 2.5 M glycine, then for a further 10 min. Samples were subsequently washed with PBS, flash-frozen and stored at −80 °C.

3.2.1.2 ChIP-seq experimental protocol

ChIP-seq experiments were carried out using the protocol described by Schmidt et al.[567]. Liver tissue samples previously harvested and cross-linked, were lysed and sonicated. DNA was immunoprecipitated for CTCF-DNA binding, and its ends repaired at 20 °C for 30 min, then we added Adenine overhang at 37 °C for 30 min,

followed by Illumina sequencing adapters ligation at room temperature for 15 min before 16 cycles of PCR amplification. PCR cycles conditions were as follows: (1) 98 °C - 30 s; (2) 98 °C - 30 s, 65 °C - 30 s, 72 °C -30 s, 16 cycles; (3) 72 °C – 5 min. Using a 2% agarose gel, DNA fragments (200 - 300 bp) were selected for 50-bp single-end read sequencing on Illumina HiSeq 2000, following the manufacturer's instructions. In order to minimise the impact of possible batch effects, biological replicates for both F0 and F1 generations were prepared and sequenced in independent flowcells.

3.2.2 Computational methods

3.2.2.1 Read mapping and measuring allele-specific binding signal

We used the *Mus musculus castaneus* genome assembly previously constructed by mapping back SNV calls from CAST on the latest version of the *Mus musculus* reference assembly (GRCm38/mm10)[257, 558]. Nucleotides at each CAST SNV position on the GRCm38 assembly were altered to represent the single variants in the other subspecies. The process mapped all SNV calls for autosomes and the X chromosome. Furthermore, a combined genome of both GRCm38.p2 (BL6) and CAST was used to map all alleles from the F1 mice.

First, we filtered and trimmed raw ChIP-seq reads using Trimmomatic (Version 0.3)[610]. Using a sliding window of 20 bp, a phred minimum score of 30 was applied and reads were retained only if they met these conditions, while having an overall length of 40 bp minimum. Next, we aligned ChIP-seq reads from the F1 mice to an alignment index of the combined genome assemblies using BWA (Version 0.7.3a)[494]. Filtered reads were aligned allowing for a maximum of two mismatches per read (-n 2), and filtering by the "XT:A:U" alignment tag, reads that aligned to multiple locations were discarded. Equivalent proportions of F1 reads mapped to the combined BL6 and CAST genomes, and to the individual genomes as well (Figure 3.1c). The ratio of BL6 to CAST CTCF binding sites in the F0 parental mice was also similar (Figure 3.1b).

Additionally, we mapped reads from F0 and F1 replicates to the GRCm38 reference genome using GSNAP[498]. Relaxed mismatch threshold criteria (allowing for a maximum of three mismatches per read) were used in order to make it possible for F1 reads of CAST origin to align back to the BL6 genome. This allowed us to assess the quality of each ChIP-seq library. Results from this analysis showed very good correspondence on the percentage of aligned reads from CAST libraries onto a BL6 background (e.g. 45% of F0 CAST libraries aligned to BL6 vs. 55% aligned to CAST genome assembly).

To determine the read coverage for each base of the joint genome assembly, we used the SAMtools mpileup program[568]. These counts were then filtered by SNV positions that allow for resolution of subspecies of origin for the DNA sequence only. A minimum of 10 reads per SNV across replicates was required for the retention of the site for further analysis. At this stage, this process was carried out at each ChIP-seq site regardless of the presence of one or more SNVs at each ChIP-seq peak. CTCF peak calling was performed using the callpeaks command from MACS2 with the parameter -g 'mm' indicating a mouse-specific effective genome size and default p-value cutoff to call peaks against input controls for both mice subspecies, BL6 and CAST, and appropriate control from their hybrid reciprocal crosses[500].

Reads were subsequently normalized for varying sequencing depth due to differences in library sizes across biological replicates in F0 and F1 mice. Normalisation for read coverage differences was done using R Bioconductor package "DESeq"[611]. This package estimates a constant scaling factor for each library/biological replicate using the median of the ratio of counts over every SNV over its geometric mean across replicates tested. The underlying assumption is that differences in read coverage at each SNV due to biological effects should only be present in a minority of sites. We then applied the resultant normalisation constant to all replicates, and these normalised read counts were used for fitting statistical models and downstream analyses.

We detected over 7,000 CTCF binding regions where more than one SNV lie in close proximity to each other, comprising about 50% of the total set of CTCF sites where an SNV with a minimum read coverage was identified and passed the various filters set earlier. However, all further downstream analysis was done using one SNV per 250 bp region in order to avoid multiple counting of the same CTCF binding sites. The overall number of reads aligned and peaks called was equivalent across replicates and generations (Figure 3.1b).

3.2.2.2 Statistical models for regulatory category assignment

CTCF binding sites were assigned different regulatory categories using ChIP-seq read counts as a proxy for the binding signal intensities of CTCF to the DNA[253] utilizing the method reported by Goncalves *et al.*[260] and Wong *et al.*[257]. "Conserved" sites were defined as those whose occupancy, despite the presence of SNVs in the binding region, between BL6 and CAST in both F0 and F1 libraries is similar. *Cis*-acting variants were defined as CTCF sites where the binding ratios between the parental BL6 and CAST F0 libraries is the same as the one obtained between the alleles in the F1 libraries, being locally determined by SNVs in the genetic sequences of these regions. CTCF binding sites influenced by *trans*-acting variants were defined

by the similarity of the signal from both alleles in the F1 despite occupancy differences between the F0 parents. *Cis–trans*-acting variation was defined as for the remainder of CTCF sites where the pattern observed comprised of both *cis* and *trans*-effects.

We modelled the counts from both BL6 and CAST parental F0 libraries as a negative binomial marginal distribution, and modelled F1 counts on a beta-binomial distribution. The parameters of the beta distribution were used to model the proportional contribution from each allele. We had 2 replicates (i) for each F0 parental subspecies and 2 replicates (j) for F1 offspring. Therefore, we assumed that F0 counts for both subspecies (x_i and y_i) to follow negative binomial distributions, whereas F1 counts (n_j) were modelled on an allele-specific basis (z_j) using a beta-binomial distribution:

$$x_i \sim \mathrm{Po}(\mu_i),\ y_i \sim \mathrm{Po}(v_i),\ z_j \sim \mathrm{Bi}(n_j, p_j)$$

$$\mu_i \sim \mathrm{Ga}\left(r, \frac{p_\mu}{1-p_\mu}\right),\ v_i \sim \mathrm{Ga}\left(r, \frac{p_v}{1-p_v}\right),\ p_j \sim \mathrm{Be}(\alpha,\beta)$$

where:

x_i = the binding intensity of the variant in the ith BL6 F0 mouse

y_i = the binding intensity of the variant in the ith CAST F0 mouse

n_j = the number of reads mapping across both allelic variants in the jth F1 hybrid

z_j = the number of reads mapping to the BL6 allele in the jth F1 hybrid.

The dispersion parameter r for F0 libraries was estimated using the function "estimateDispersions" from the "DESeq" Bioconductor package with local regression fit. r was defined as the reciprocal of the fitted dispersion value as computed using "estimateDispersions". Parameter estimation for the two distributions was constrained based on the four different regulatory scenarios outlined above, and we derived maximum likelihood values for all four scenarios on a "site-by-site" basis, as follows:

Conserved: $p_\mu = p_v$ and $\alpha = \beta$

Cis: $p_\mu \neq p_v$ and $\frac{\alpha}{\alpha+\beta} = \frac{\frac{p_\mu}{1-p_\mu}}{\frac{p_\mu}{1-p_\mu}+\frac{p_v}{1-p_v}}$

Trans: $p_\mu \neq p_v$ and $\alpha = \beta$

Cis-trans: $p_\mu \neq p_v$ and $\alpha \neq \beta$.

The most likely model at each site was determined using the Bayesian Information Criteria (BIC) estimation for each of the hypothetical cases. stated earlier, only variants separated by a minimum distance of 250 bp were used for this and all other analysis.

To allow for meaningful comparison, the statistical model described above was also used to assign the four regulatory categories in two biological replicates, selected randomly from the set of 6 biological replicates reported in Wong et al.[257] for the three liver-specific transcription factors (TFs): CEBPA, FOXA1 and HNF4A. Raw reads from these two replicates were normalized for differences in library sizes across replicates in F0 and F1 mice, and their dispersion parameter r was estimated similar to the method discussed above prior to be fitted with the same statistical models. An example of the R code used for statistical modelling and category assignment is available in the appendix.

3.2.2.3 Subsampling strategy to investigate the effect of library availability on regulatory category assignment

Due to the availability of only two biological replicates for the analysis of *cis*/*trans* variation effect on CTCF occupancy in hybrid F1 mice, a subsampling strategy was undertaken to evaluate the effect of extra libraries inclusion on the results of regulatory category assignment. The approach consisted of randomly selecting 2 biological replicates from a possible 6 replicates for each of the F0 parental subspecies and their reciprocal F1 hybrid progeny (ensuring that the for each F1 replicate the reads for both alleles corresponding to the same animal are included) for each of the liver-specific TFs mentioned above in section 3.2.2.2. The raw reads overlapping the full set of SNVs were retrieved, normalised across all replicates in F0 and F1 for library size differences, and had their dispersion parameter calculated as previously detailed. Normalised reads in those replicates were then fitted with the statistical models to assign regulatory category based on the binding site signal intensities peculiar to those replicates. An example of the R code used for statistical modelling and category assignment is available in the appendix.

This was repeated for 1000 random combinations of any two replicates for any of the 6 for each F0 and F1 line, for each TF separately. The subsampling strategy was then repeated for an increasing number of replicates, running it for 1000 runs of random combination of 3, 4 and 5 replicates for each TF. The estimations of regulatory category from each run, for each number of replicates and each TF were outputted and used to estimate the effect of incorporating extra biological replicates on the accuracy of category assignment.

We compared this effect to the category assignment proportions derived from the full set of 6 biological replicates available for each TF. The CTCF estimate was calculated from 2 replicates only by multiplying the fraction of each category by the total number of SNVs in each TF to generate TF-specific estimates for each category.

3.2.2.4 CTCF inter-peak coordination of binding intensity

Correlation coefficients for the binding signal intensities between pairs of CTCF sites at incremental genomic intervals were calculated in order to test whether these genomic regions are under the influence of CTCF *cis*-acting regulatory variants. Spearman's correlation coefficient of allelic proportions (BL6/(BL6+CAST)) was computed between CTCF sites at successive bins anchored by *cis*-acting variants to capture the coordination between CTCF sites at anchor points and those in the consecutive bins. Spearman's ρ for each mutually exclusive bin with the corresponding anchor CTCF site was determined, increasing the interval to the next bin by one extra kb (1 kb) from the *cis*-acting variant.

To test the decay in signal by increasing distance, Spearman's ρ estimates for the entire set of CTCF *cis* sites at each distance were used as the outcome variable in a linear regression model using log-transformed distances as the predictor variable. A null distribution for the correlation of binding signal was created by comparing binding levels of anchor CTCF sites with the other CTCF locations randomly sampled across the genome. We subsampled an equal number of CTCF sites randomly from the total pool without replacement as null anchor points, then simulated a set of binned peaks for each null anchor, keeping them constant. The total number of these null anchor peaks and their simulated binned peaks pair was equal to the total number of anchored–binned peak pairings originally observed. Spearman's ρ estimates for this null set was then calculated similarly, and the values were fitted with a linear regression model.

3.2.2.5 Statistical models for lineage-specific CTCF occupancy

We used a similar statistical model to tease apart the influences of *cis* and *cis*–*trans* acting variants on lineage-specific CTCF binding sites. Normalized read counts between F0 and F1 libraries were used to define lineage-specific CTCF sites based on the ratios of F0 and F1 ($\text{ratio}_{F0} = B6_{F0}/(B6_{F0}/CAST_{F0})$ and $\text{ratio}_{F1} = B6_{F1}/(B6_{F1}/CAST_{F1})$), where ratios were calculated between mean levels of binding signal across biological replicates. Lineage-specific sites were defined using the following criteria: ($\text{ratio}_{F0} < 0.05$ and $\text{ratio}_{F1} < 0.05$) or ($\text{ratio}_{F0} > 0.95$ and $\text{ratio}_{F1} > 0.95$).

A lineage-specific site solely influenced by *cis*-acting variation would have F1 read counts that are half of that in F0. Variants acting in *trans* would cause a significant stray from this 2:1 ratio. To test the likelihood of these two scenarios, a statistical model was modelled using the negative binomial distribution, and applied

to each lineage-specific site, followed by maximum likelihood estimation and BIC to select the best fitted model.

$$x_i \sim \mathrm{Po}(p_i),\ 2y_i \sim \mathrm{Po}(o_i)$$

$$p_i \sim \mathrm{Ga}\left(r, \frac{S_{\mathrm{pmax}}}{1-S_{\mathrm{pmax}}}\right),\ o_i \sim \mathrm{Ga}\left(r, \frac{S_{\mathrm{o}}}{1-S_{\mathrm{o}}}\right),$$

where:

x_i = the normalized read count binding intensity of the variant in the ith F0 mouse from the subspecies of lineage-specific binding

y_i = the binding intensity of the variant in the ith F1 mouse summed across both alleles.

The dispersion parameter, r, was estimated using "DESeq". The two following scenarios were tested:

$$\textit{Cis}\text{: } S_{\mathrm{pmax}}=S_{\mathrm{o}}$$
$$\textit{Cis-trans}\text{: } S_{\mathrm{pmax}}\neq S_{\mathrm{o}}.$$

Results from the lineage-specific statistical modelling outlined above were compared to estimates derived from fitting 2 randomly-selected replicates for the three liver-specific TFs mentioned above with the same models to facilitate comparison. An example of the R code used for statistical modelling and category assignment is available in the appendix.

3.2.2.6 Statistical models for CTCF inheritance mode assignment

Normalized read counts across all F0 and F1 libraries were used to investigate the mode of inheritance of CTCF binding intensities at genomic locations characterised by *cis*- and *trans*-acting variation. The counts from every SNV were then fitted to statistical models assessing either additive or dominant/recessive inheritance modes. The statistical models were constructed based on the assumption that if F1 binding intensities were inherited in an additive manner, the total binding intensity from both alleles should, in theory, equal the total binding intensity of F0 summed across both parental alleles. If these were, however, showing a dominant/recessive mode of inheritance, the total binding intensity from both alleles should equal the total binding intensity of one of the F0 parents and not the other. The inheritance modes were modelled on negative binomial distributions defined as follows:

$$x_{\mathrm{max},i} \sim \mathrm{Po}(p_{\mathrm{max},i}),\ x_{\mathrm{min},i} \sim \mathrm{Po}(p_{\mathrm{min},i}),\ y_i \sim \mathrm{Po}(o_i),$$

where:

$x_{\max,i}$ = the normalized read count of the variant in the ith F0 mouse showing the higher median binding intensity among replicates

$x_{\min,i}$ = the normalized read count of the variant in the ith F0 mouse with the lower median binding intensity among replicates.

y_i = the binding intensity of the variant in the ith F1 mouse summed across both alleles.

$$p_{\max,i} \sim \mathrm{Ga}\left(r, \frac{S_{\mathrm{pmax}}}{1-S_{\mathrm{pmax}}}\right),\ p_{\min,i} \sim \mathrm{Ga}\left(r, \frac{S_{\mathrm{pmin}}}{1-S_{\mathrm{pmin}}}\right),\ o_i \sim \mathrm{Ga}\left(r, \frac{S_{\mathrm{o}}}{1-S_{\mathrm{o}}}\right).$$

As done previously, the dispersion parameter, r, was estimated using "DESeq". Maximum likelihood estimation was used to fit the counts, followed by BIC to assess which of the following two models best fit the binding intensity from each CTCF site affected by variation in *cis* or *trans*.

Dominant: $S_{\mathrm{pmax}}=S_{\mathrm{o}}$ or $S_{\mathrm{pmin}}=S_{\mathrm{o}}$
Additive: $S_{\mathrm{pmax}}\neq S_{\mathrm{o}}$ and $S_{\mathrm{pmin}}\neq S_{\mathrm{o}}$.

Sites where the parameter estimated for the offspring, S_o, could not be resolved from the parameters estimated for both parents (i.e., if $S_o = S_{pmax}$ and $S_o = S_{pmin}$) were excluded from the results. These sites were identified by testing both modes of inheritance separately for $p_{max,i}$ and $p_{min,i}$, and discarding sites that fit the dominant model in both cases. An example of the R code used for statistical modelling and category assignment is available in the appendix.

Only sites assigned to each mode with a BIC > 1 were used in the analysis to further improve the accuracy of the model. The stringent criteria rendered the use of this model to assign modes of inheritance in 2 randomly-selected replicates in the liver-specific TFs impractical, as very few *trans*-acting variants were found to have a BIC >1. We, therefore used the full set of 6 biological replicates to assign the mode of inheritance using the model described above and used the estimates derived for comparison with CTCF.

3.2.2.7 Cross-Tissue Analysis of *cis*/*trans*-acting variation in CTCF binding

The same CTCF ChIP-seq libraries for adult BL6 male mice used for the equivalent analysis in section 2.2.2.5, were retrieved from the ENCODE Project data repository[451]. The analysis compared *cis*/*trans*-influenced CTCF occupancy in 12 tissues: lung, bone marrow, bone marrow macrophages, cortical plate, cerebellum, heart, kidney, thymus, spleen, olfactory bulb, small intestine and testis. using BEDTools intersect 2.2.5.0 with the options -wa -wb, the overlap between *cis*/*trans*

liver CTCF peaks and the peaks from every ENCODE tissue was used to investigate the tissue-sharedness of these binding sites. UpSet plots were generated using the ComplexHeatmap package in R[571].

Similar to analysis explained in section 2.2.2.5, we calculated the tissue diversity index for these *cis*/*trans* CTCF sites using the log10 of the p-value at peak summit computed by MACS during peak calling. P-values for each *cis*/*trans* CTCF binding site were used to calculate the Shannon Diversity Index for each tissue using Vegan package in R[572]. CTCF binding conservation across tissues was measured by the proportion of CTCF binding sites bound within each bin of Shannon diversity index.

The tissue conservation analysis was repeated for sites defined as lineage-specific in 3.2.2.4. Subsequently, the extent of tissue-wide occupancy conservation was investigated across the categories of *cis*/*trans* variants in CTCF binding sites for both the total set and lineage-specific set.

3.2.2.8 Analysis of the effect of incorporating extra biological replicates on *cis*/*trans* variation

Increasing number of replicates were randomly selected from the full set of 6 for each of the three TFs, CEBPA, FOXA1 and HNF4A and used to analyse the effect of increasing number of replicates on the various aspects of *cis*/*trans* regulatory effects on TF occupancy detailed above. After running the normalisation followed by statistical modelling for regulatory category assignment in section 3.2.2.2, the difference between the smallest and second smallest BIC values (dif_BIC) was used as an estimate of the reliability of *cis*/*trans* variant calling. The higher the value of BIC, the more reliable the call was. We used a minimum BIC value of >=1 to analyse the effects of increasing replicate number on the type and number of high-confidence category assignment.

We additionally used the random set of 2-5 replicates for each TF to estimate the effect of *cis*/*trans* variation on their occupancy. Pearson's correlation coefficient (r) was measured between ratios of BL6 in F1 vs F0. Correlation coefficient of 0 indicated equal effect of *cis*- and *trans*-acting variations, whereas a correlation coefficient of 1 suggested the lack of any *trans* influences. The results were compared to the *cis*/*trans*-acting variation effect on occupancy for the full set of 6 replicates.

Furthermore, the random set of 2-5 replicates were analysed for the effect of replicate number on the type of lineage-specific binding in each TF. Of particular interest was the how the addition of more replicates enhances the ability to define

lineage-specific binding, and tease apart the effect of *cistrans* variants on F1 allelic binding intensity (diversifying vs. compensatory). Modes of inheritance for increasing number of replicates were also determined using the statistical models described in 3.2.2.6 in *cis* and *trans* variants of each TF. Analysis of inheritance mode was restricted to sites where the absolute difference in the average binding intensities across the F0 parental subspecies was greater than or equal to twice the standard deviation of the average binding signal across biological replicates (>=19 normalised read). To further minimise the noise from the data, only sites whose *cis*/*trans* variation was assigned with a minimum of BIC >=1 were used to improve the reliability of the results.

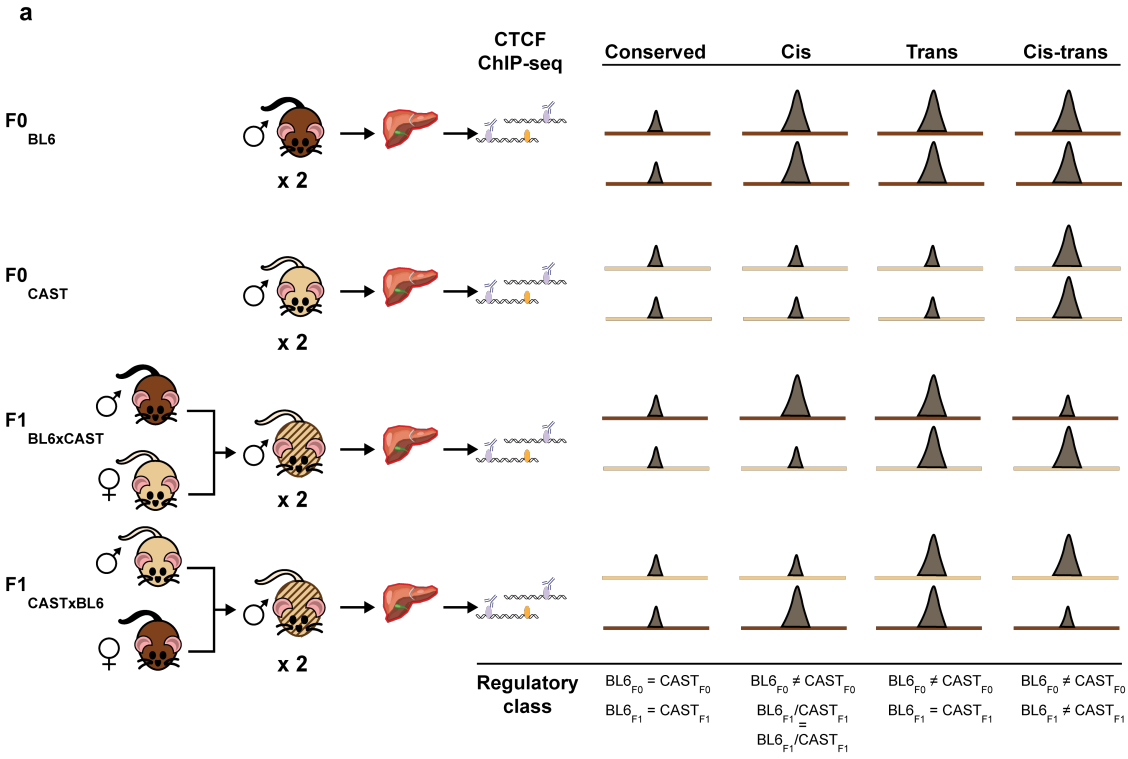
3.3 Results

3.3.1 Equal *cis* and *cistrans* effects on CTCF occupancy

CTCF-bound genomic regions were retrieved from liver samples of adult mice of the same two inbred mice subspecies of the *Mus musculus* genus used for the analysis in Chapter 2: *Mus musculus domesticus* (C57BL/6J or BL6 for short) and *Mus musculus castaneus* (CAST), and their F1 hybrid offspring of two reciprocal crosses (BL6xCAST and CASTxBL6). Binding sites were derived from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) libraries of CTCF in two biological replicates for each of the four lines, for a total of 8 biological replicates (2 for each F0 parental subspecies and 2 for each of the two F1 hybrid crosses) (Figure 3.1a). Differences in binding affinities, by proxy of varying read enrichments, between the F0 parental mice and their F1 hybrid offspring were used to discern the evolutionary, genomic and functional dynamics of *cis* and *trans* variation and their impact on the pattern of CTCF binding.

By using normalised ChIP-seq read counts, different regulatory categories were assigned to each binding sites, based on the presence of single nucleotide variants (SNVs) overlapping ChIP-seq peaks and the differences in their read enrichments. Using the approach detailed in the Methodology, we were able to distinguish between four possible regulatory categories that describe the binding of CTCF between the two subspecies into 4 categories: *conserved*, *cis*, *trans*, *cistrans* (Figure 3.1a). Under the *conserved* category, SNVs do not exhibit any measurable differences in their binding signal intensities between either F0 parental alleles or their F1 progeny. *Cis*-acting variant read enrichment is associated with that of the specific parental allele [257, 260, 612]. Despite distinctly different read signals from both F0 parental alleles, *trans*-acting variation affect both alleles in F1 equally, due to diffusible elements in the shared nuclear environment. Lastly, the *cistrans* classification encompasses the remainder of cases where allelic binding intensities were different between both

parental subspecies, but their binding among their progeny was also observed to be different. This could indicates either a combination of *cis*- and *trans*- acting variation influencing CTCF binding in F1 due to both a common environment and allele-specific effects, or insufficient signal to allow for confident category assignment that may improve with adding more biological replicates.



b

Strain	Library ID	# Reads	# Aligned	% Aligned	# Peaks
BL6	do3066	36285855	21225921	58.50%	46685
	do3073	31507903	20075616	63.72%	47183
CAST	do3069	42127765	21580214	51.23%	41260
	do3075	43617989	27179619	62.31%	39825
BL6xCAST to BL6	do3068	37449877	20440930	54.58%	46631
	do3074	37991636	20768742	54.67%	52220
CASTxBL6 to BL6	do3077	42033601	23864049	56.77%	42865
	do3078	35797333	18395785	51.39%	40435

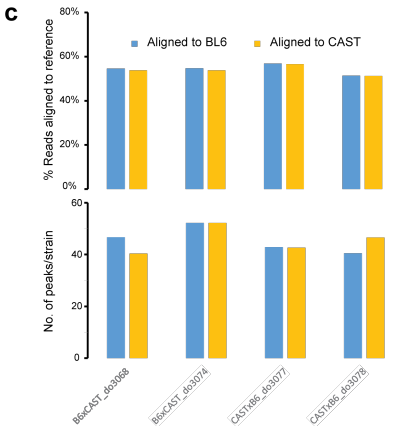


Figure 3.1: Overview of the experimental design and preliminary results.

a CTCF occupancy was profiled using ChIP-seq of liver samples of male mice from C57BL/6J (BL6), CAST/EiJ (CAST), and their reciprocal F1 crosses: BL6xCAST and CASTxBL6 in 2 biological replicates for each genetic

background. Normalized ChIP-seq read counts between BL6 and CAST at SNVs were used to assign regulatory classes for variation in CTCF binding. Based on the schematic diagram, by comparing BL6 and CAST ratios between F0 and F1 mice, CTCF binding sites can be classified into four regulatory categories: conserved, *cis*, *trans*, and *cistrans*. **b** Summary table for the results of library alignment and peak calling sorted by genetic background. Lower mapability of aligned reads is caused by the stringent criteria of maximum mismatch of two bases per read (See Methods). **c** Bar plots of the results of ChIP-seq read alignment of F1 libraries to both F0 genomes. Top plot shows the F1 reads aligning to BL6 and CAST genomes, whereas the bottom plot shows F1 CTCF peaks called from reads aligned to BL6/CAST genomes (in 1000s). Similar number of peaks were called in F1s with BL6 or CAST genomes the majority of them overlap over 90% reciprocally.

A sufficient read depth was obtained following ChIP enrichment, and >20 million reads (over 50% of all reads) from each replicate aligned to their corresponding genome despite stringent mapping criteria (See Methods) (Figure 3.1b). A comparable number of CTCF peaks was obtained from all replicates in both F0 and F1, with a mean of > 44,000 peaks per replicate, consistent with the overall number of CTCF binding sites previously reported[269, 559]. Notably, an equal proportion of F1 ChIP reads mapped back to both parental, BL6 and CAST, genomes, and produced a comparable number of binding sites when each set of aligned reads were peak called (Figure 3.1c), confirming the ability to map back alleles from each F1 hybrid to their parent of origin, and the feasibility of drawing comparisons between the two subspecies.

A total of over 58,000 CTCF binding sites were identified across replicates/crosses. The majority of these sites (75% of all binding sites) were not characterised by the presence of SNVs, thus it was not possible to investigate their allelic differences in binding between the two subspecies as they could not be told apart. There were, however, about 25% of binding sites that had one or more SNV within the peak region with sufficient read enrichment signal to quantitively resolve the difference in allelic binding in both the F0 and F1 mice (Figure 3.2a). Half of these sites have a single SNV in the peak region, with the remainder carrying two or more SNVs in their sequence. The numbers obtained in this analysis roughly compare to those obtained in a study looking at the other liver-specific transcription factors (Figure 3.2a)[257]. In order to avoid conflating the results by repeatedly counting binding sites with more than one SNV, all downstream analysis used SNVs that are at least 250 bp from the next SNV, restricting them to a single SNV per binding site (see Method).

CTCF binding sites with SNVs informative for allelic differentiation were assigned one of the four classes outlined above using statistical modelling (see Methods).

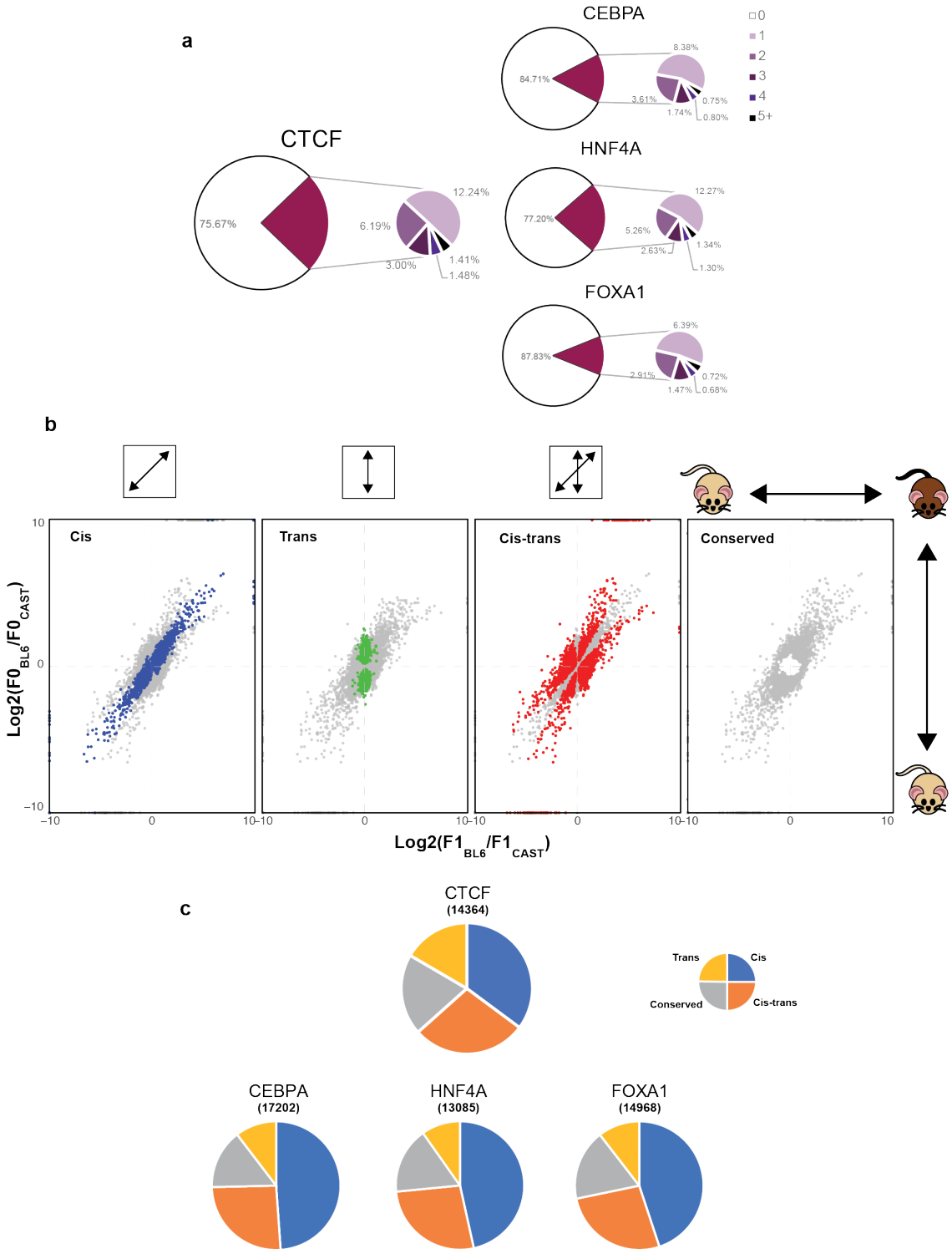


Figure 3.2: Regulatory categories assignment demonstrates that CTCF occupancy levels are equally *cis*- and *cistrans*-driven for 2/3 of sites.

a Pie charts for the number of CTCF binding sites obtained after peak-calling and SNV mapping with associated number of SNVs. The percentages for the nested pie charts reflect their proportions out of the total number of binding sites. **b** Scatterplots of BL6 vs CAST log2 ratios of CTCF binding intensity signals in F0 and F1 mice. Every point represents an SNV. Regulatory categories-assignments are colour-highlighted in individual scatterplots. Direction of distribution of SNVs is indicated above each plot. Grey-coloured points are the remainder of SNVs that are not assigned to the highlighted category. **c** Pie charts displaying the regulatory class make-up of SNVs overlapping CTCF compared to those derived from 2 randomly selected replicates for each of the three other TFs. The numbers in brackets indicate the total number of TF binding sites with a minimum of one SNV in their sequence for each TF.

In order to verify this class assignment, the differences in the binding ratio between the F0 and F1 alleles were visualised as the ratios of the F1 BL6 allele to its CAST counterpart against the corresponding ratio of F0 alleles. As seen in Figure 3.2b, *Cis*-acting variants cluster along the diagonal line as their F1 ratios correspond to those of the parental lines, whereas *trans*-acting variants form a straight line parallel to the F0 ratios, a result of their departure from their parental alleles signals. *Cistrans*-variants significantly deviate from the diagonal line, filling the area between *cis*- and *trans*- variants (Figure 3.2b).

In total, there were 14,364 CTCF binding sites characterized by the presence of SNVs with sufficient read coverage to allow the resolution of allelic difference, and that we were also able to assign regulatory categories to, equivalent to the number of sites used for other TFs (13,000 – 17,000) (Figure 3.2c). Although CTCF binding sites, similar to liver-specific transcription factors, were most frequently influenced by *cis*-acting SNVs (35%), these were followed very closely by *cistrans* (28%), then conserved (20%) and *trans* variants (17%) (Figure 3.2c). The enrichment of *cis* variants on CTCF was noticeably lower than observed in the liver-specific TFs with an equal number of biological replicates. CTCF, on the other hand, showed a marked increase in the fraction of *trans* regulatory variation in occupancy compared to all other three TFs (17% vs 10%). Estimates for the contribution of conserved (*cons*) variation in CTCF and other TFs binding were equivalent (Figure 3.2c). Assignment of binding regulatory variation in CTCF was found to be statistically significantly different from all 3 liver-specific TFs (χ^2 test for pairwise comparison between CTCF and other TFs with Bonferroni correction, all p-values $<$ 2.2e-16).

Although *cis*-acting variation was the most prominent mode of variation present in the liver-specific TF binding sites in the original analysis that used 6 biological replicates, conserved binding, even when associated with sequence changes in the form

of SNVs, is always the second most common type of regulatory variation in TF occupancy, with *trans* and *cistrans* variation contributing less than third[257]. The increase in the *cistrans* effect size when the analysis was run in two biological replicates instead seems to have come at the expense of both the *conserved* and *cis* variants (Figure 3.2c).

A subsampling approach was undertaken in order to assess the validity of the assignment of *cis*/*trans* categories to CTCF binding sites based on only two biological replicates and ensure comparability with the three other liver-specific TFs. The aim was to elucidate the effect of biological replicate number on the ability to resolve the difference in read coverage into distinct regulatory categories. By randomly combining a specific number of biological replicates (from 2 to 5) for each of the three TFs, then running the *cis*/*trans* category assignment algorithm for one thousand times, we were able to obtain the range of category estimates for each of the four subsampling strategies (Figure 3.3). The results of the subsampling strategy of biological replicates in other TFs show an overall improvement in the estimates of the four regulatory categories with the addition of every extra replicate towards the values obtained when the experiments where run with 6 biological replicates (Figure 3.3 density plots). This is additionally evidenced by the narrower distribution of values, reflecting less dispersion of values, with increasing replicate number (Figure 3.3 dot plots).

These improvements, however, are not uniformly distributed among the *cis*/*trans* categories. The most conspicuous change is invariably observed with the resolution of *cistrans* sites into other categories, as their proportions strongly decrease with the addition of extra replicates (starting from 3 replicates). The range of value obtained in all *cistrans* 2-replicate runs for all TFs never matches the original 6-replicate estimate. Conserved sites (*cons*) show exactly the opposite pattern, increasing considerably with the addition of extra replicates. The range of *cons* values for 2 replicates is not that of 6-replicate. The estimates of *cis*-influenced do slightly increase with increasing the number of replicates, but the overall distribution of values for 2-replicate runs mostly overlaps with that of higher replicate number, and occasionally (especially in the case of CEBPA) considerably overlaps with the 6-replicate estimate. The pattern for *trans* sites is even subtler, with tighter ranges of values, and estimates that do not generally deviate from the 6-replicate estimate. For example, CEBPA 2-replicate mean values are nearly at the 6-replicate estimate (Figure 3.3).

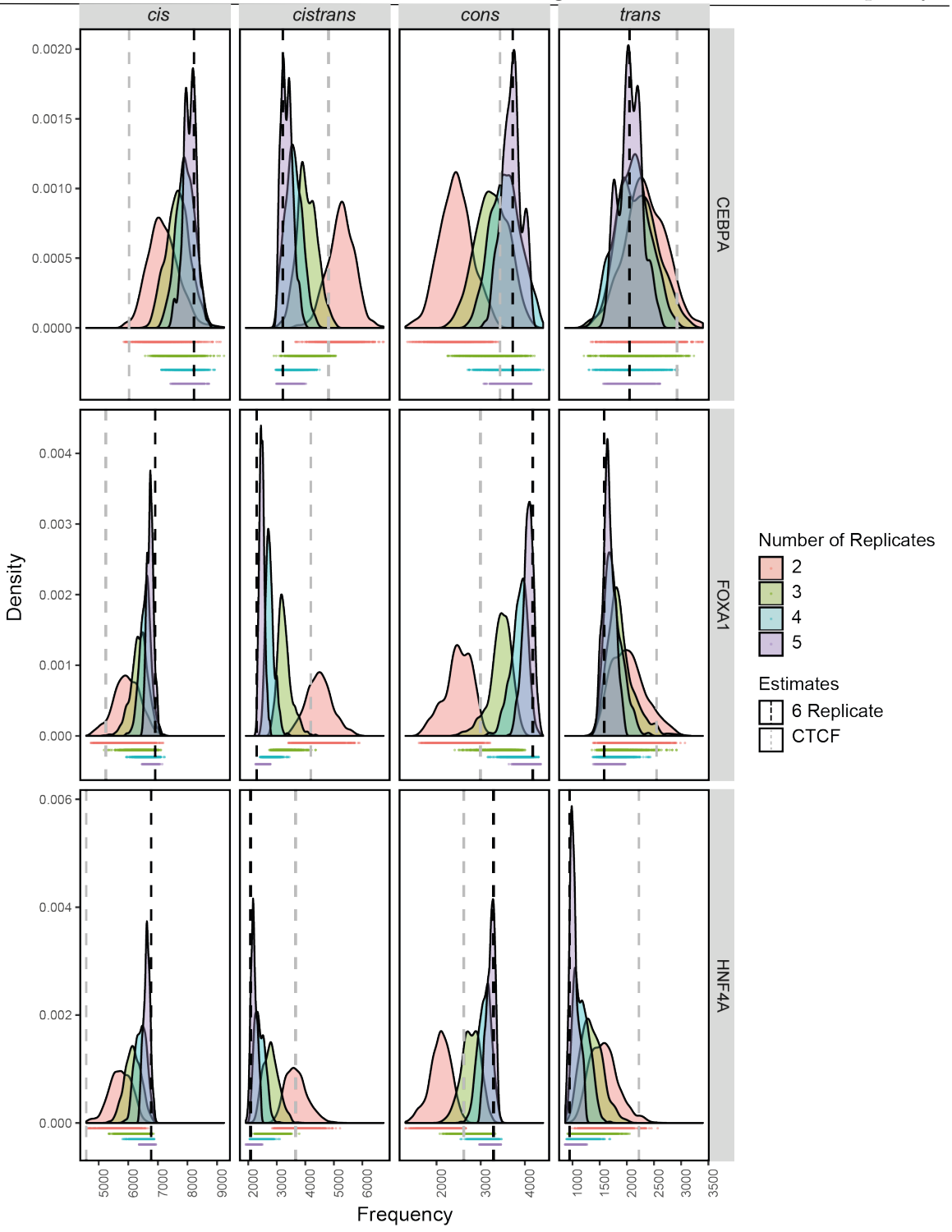


Figure 3.3: Ascending subsampling of biological replicates in other TFs support the *cis* and *trans* proportions observed in CTCF.

Density plots of all 1000 randomised combinations of biological replicates in ascending number of replicates (from 2-5) for the 3 liver-specific transcription

factors (CEBPA, FOXA1 and HNF4A), faceted by the 4 regulatory categories: *cis*, *cistrans*, conserved (*cons*) and *trans*. The area under each curve correspond to the entire range of values (number of binding sites classified as such after one run of the algorithm, repeated for a 1000 randomised runs) for each category, number of replicates and TF. The horizontal dot plots under each facet illustrate the distribution of the values obtained for each category per number of replicates in that TF, and correspond to the width of the curve above. The black dashed line indicates the original estimate for the number of sites for the particular category in that TF as derived from the original analysis in Wong *et al.*[257]. The grey dashed line indicates the number of CTCF sites calculated for 2 biological replicates, estimated for every TF based on the proportion of each particular category in CTCF, and multiplied by the overall number of sites in each TF.

These results validates the proportions of the *cis* and *trans* observed in CTCF. Although the *cistrans* estimate for CTCF almost always overlaps the mean/median for 2-replicate runs in other TFs, and may similarly resolve into other categories with the addition of extra biological replicates for CTCF, the estimates for the three other categories appear different than those of the other TFs (Figure 3.3). The estimate for *cons* sites were generally higher than the equivalent 2-replicate mean values (almost equal to mean/median of 3-replicate runs in CEBPA and HNF4A). CTCF *cis* variants estimates are always lower than any of their 2-replicate counterparts in other TFs. Although this estimate may similarly go up with the addition of extra replicates (via the resolution of *cistrans* sites), on this evidence CTCF *cis* variants would remain less abundant than in other TFs. Conversely, the CTCF *trans* estimate is always much higher, and as *trans*-assigned sites only slightly decrease with added replicates, the CTCF *trans* component looks to be distinctly higher.

Even though *cis*-acting variation was the most common in CTCF binding, the effect size on its occupancy was also different. This is clearly reflected in the Pearson's correlation coefficient of CTCF occupancy between the two parental subspecies and their offspring (Figure 3.4a). When *cis*- and *trans*-acting variations have an equal effect on occupancy differences between the F0 and F1, absence correlation (correlation coefficient = 0) would be observed, whereas a perfect correlation (correlation coefficient = 1) signals the lack of any *trans* influences. Although all four TFs (CTCF, CEBPA, FOXA1 and HNF4A) showed correlation coefficients that are considerably large (r >= 0.7, all p-values $<$ 2.2e−16), the distribution of read enrichments from CTCF binding sites exhibit a higher degree of dispersion and deviation from the strongly-*cis* pattern seen in other TFs (Figure 3.4a). This is further evidenced by a lower correlation coefficient ($r = 0.70$), that is statistically significantly different compared to that for CEBPA, FOXA1 and HNF4A (z-test, all p-values < 0.0001).

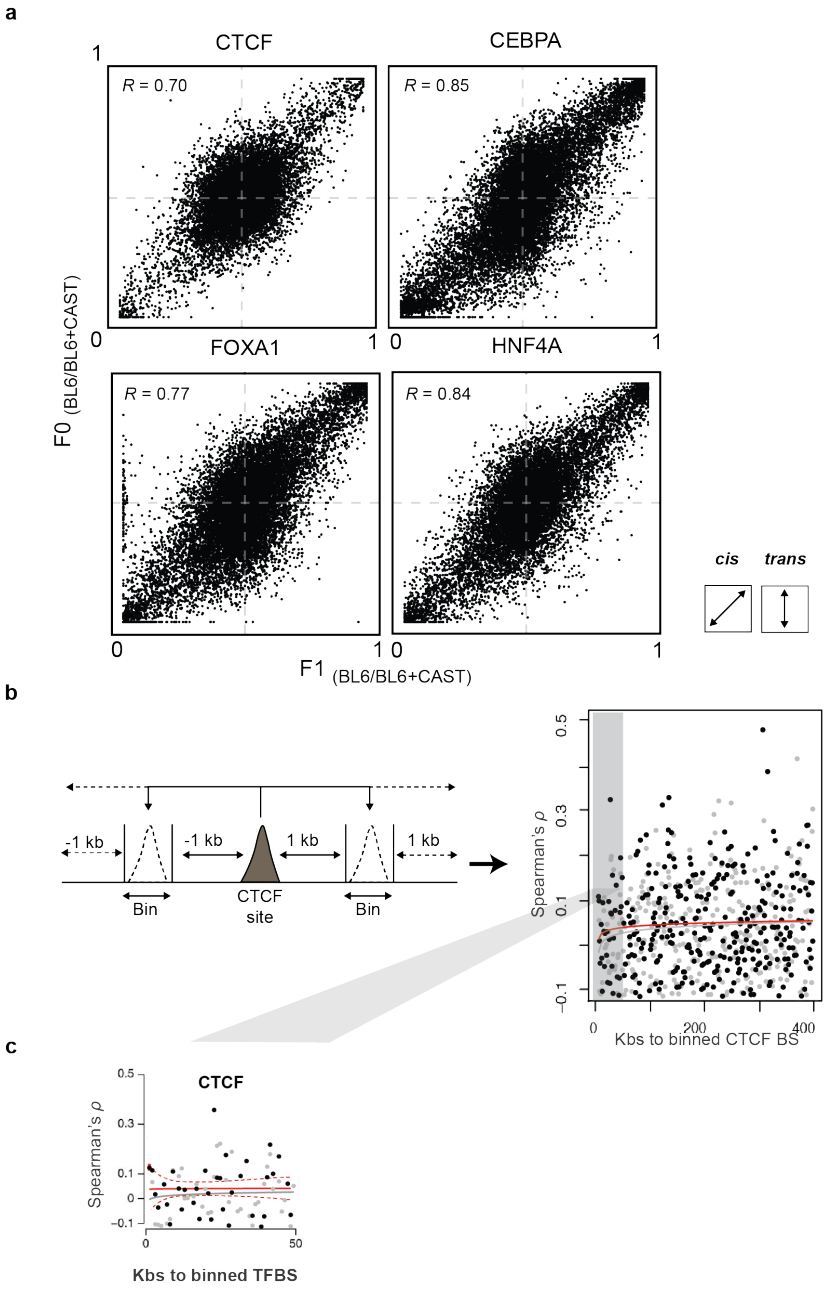


Figure 3.4: *Cis*-acting variants do not display inter-peak correspondence in CTCF.

a Scatterplots for the mean F0 vs. F1 binding intensity ratios (BL6 vs. CAST) for CTCF (left) and the 3 liver-specific TFs (right). Data from 2 randomly selected replicates for the other TFs were used for plotting to allow for meaningful comparison. The correlation coefficient (r) indicate the level of *cis*-directed regulatory effect. **b** Schematic diagram (left) for the method of measuring the span of *cis*-regulatory effects. Consecutive 1 kb bins were taken

from each CTCF site affected by *cis*-acting variation in both directions for a distance of 400 kb. Spearman's ρ was computed for each bin through the BL6:CAST allelic ratio between SNVs in bins vs anchored SNVs. Spearman's ρ values for each bin (right) were plotted (black dots). Red line is the linear regression line. Grey dots represent the null distribution of random subsampling from the total set of *cis*/*trans* CTCF sites. The grey line is the linear regression line for the Spearman's ρ values from the null distribution. **c** A blow-up of the Spearman's ρ values for each bin in the 50 kb range from the anchorage points for CTCF. The red line is the linear regression line and the red dashed lines mark the 90% confidence intervals of the slope of the line. Grey dots represent the null distribution. The grey line is the linear regression line for the Spearman's ρ values from the null distribution.

3.3.2 Effect of distance on *cis*-acting inter-peak correspondence

Whereas *cis* variation effect size decreases at a logarithmic rate with increasing genomic distance in other TFs with tissue-specific activity, CTCF *cis* acting variants do not seem to exert any effect on either proximal or distal CTCF binding sites. The correspondence of ratios between *cis* and other SNVs at variable genomic distance cannot be distinguished from the null distribution of the genomic background, and the linear regression line is flat. The effect is absent in both proximal (<50 kb) and distal (400 kb) distances (Figure 3.4b, c). This indicates that the genomic scope of *cis*-acting variants, proposed to be of short-range[257, 613], does not necessarily hold true for CTCF. CTCF binding intensity between proximal sites is not dependent on the presence of other nearby CTCF sites (*cis* or otherwise), hence they do not show the pattern of decay of signal correlation with increasing distance between their genomic positions.

3.3.3 Lineage-specific CTCF binding is driven by *cis* variation

We next applied statistical models to study *cis*/*trans* dynamics in CTCF and tissue-specific TFs occupancy, to test *cis* and *cistrans* effects in lineage-specific CTCF binding (see Methods). We tested whether such variation caused by SNVs on CTCF binding could be responsible for the rise of novel binding sites in a lineage-specific fashion. Lineage-specific binding was defined as occupancy events detected in one parental F0 subspecies, with a single allele-specific corresponding signal in F1 (Figure 3.5a). If the divergence is the product of *cis* acting variation only, the binding signal intensity in the corresponding F1 allele will be half that of the parental subspecies, whereas if

CTCF binding is additionally influenced by *cistrans* variation, the binding intensity of the signal in the F1 allele would be either stronger or weaker than half of that observed in the F0 subspecies of origin (Figure 3.5a).

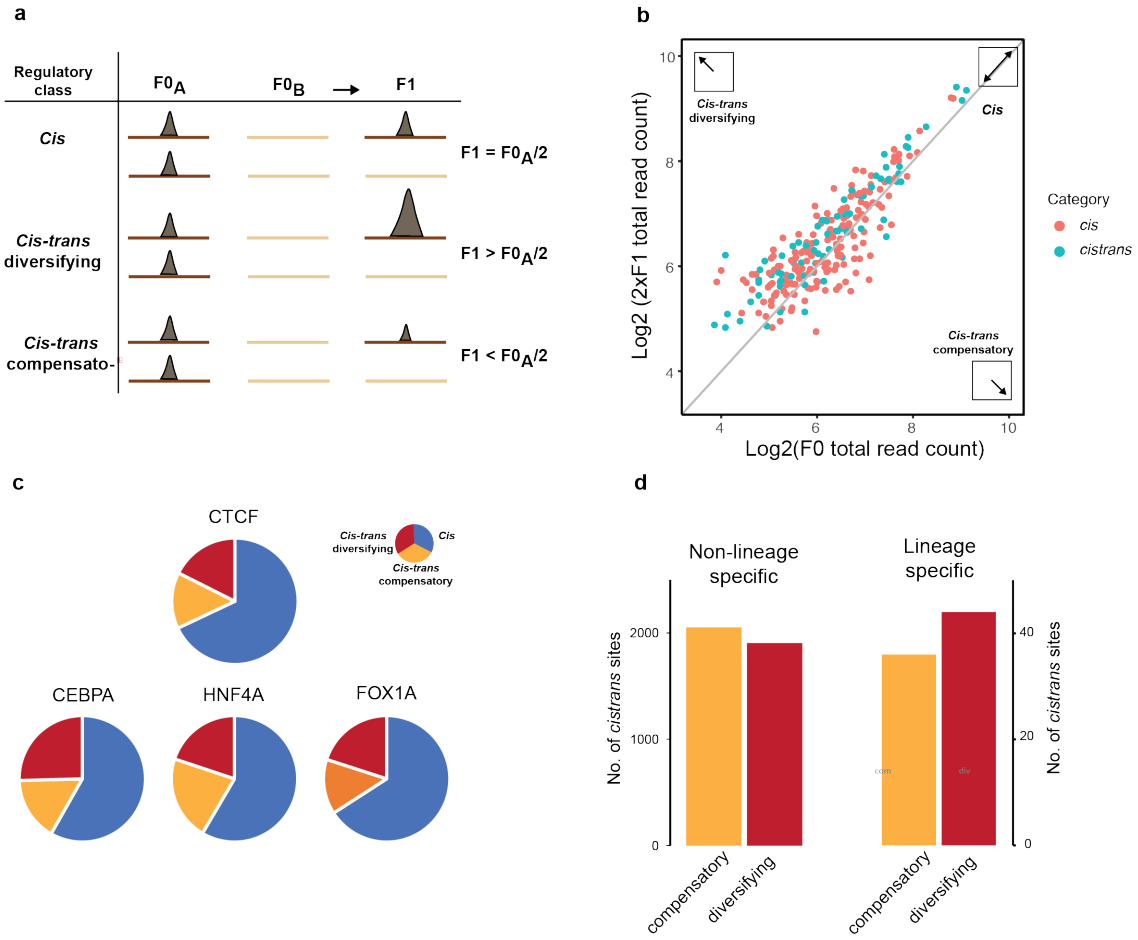


Figure 3.5: Lineage-specific CTCF occupancy is driven by *cis*-acting variation

a A model for lineage-specific class assignment. Lineage-specific binding sites were those where binding occurs exclusively in either BL6 or CAST in F0 parents and in an allele-specific manner in F1 individuals based on a cut-off ($F0_{B6/(B6+CAST)} > 0.95$, $F1_{B6/(B6+CAST)} > 0.95$, $F0_{B6/(B6+CAST)} < 0.05$, $F1_{B6/(B6+CAST)} < 0.05$), sorted into three categories. **b**. A scatter plot of average CTCF log2 F0 total read counts against average log2 F1 read count (BL6 + CAST allele) multiplied by 2, using averages across biological replicates. CTCF binding sites affected by *cis*-acting variants are expected to distribute along the diagonal. CTCF binding sites affected by *cistrans*-acting variants will deviate from the diagonal. The direction of this deviation is an indication of the type of lineage-specific variation described in **a**. **c** Pie charts showing the relative proportions of the three categories set in **a** for CTCF and 2 replicates for the 3 liver-specific TFs estimated using a

statistical model to fit binding sites affected by *cis* or *cistrans* variation (see Methods). **d** Bar plots showing compensatory vs diversifying *cistrans* modes of regulation. These are compared here to *cistrans* sites from non-lineage-specific CTCF binding sites.

Cistrans acting variants influencing binding of CTCF in lineage-specific sites exert these effects in two modes reflecting different selective forces. Those binding sites that exhibit either greater or lower signal intensity in the F1 allele compared to the signal from the corresponding F0 can be classified into *cis*-acting variants that are either *compensated* or *diversified* by *trans*-acting variation (i.e. *cistrans*). If the binding intensities in the F1 are lower than in the F0, these changes were classified as *compensatory*, whilst they were deemed *diversifying* if those binding intensities are greater in the F1 than in the F0 (Figure 3.5b). Under the null hypothesis, the effects from *trans*-influenced variation on lineage-specific binding should not substantially be preferential towards either selective force. There is almost equal contribution from compensatory *cistrans*-acting variation on the lineage-specific binding of CTCF (36/250) to diversifying *cistrans* variation (44/250) (Figure 3.5b). The effect seen here in CTCF is equivalent to previously observed in liver-specific TFs (Figure 3.5c).

Based on the statistical model classification of lineage-specific binding events, over two thirds (68%, 170/250) of these CTCF sites were assessed to be under the influence of *cis* variants exclusively. The remainder of sites (32%, 80/250) were influenced by variation acting in *cistrans* (Figure 3.5c). There is; however, a major difference in terms of the number of lineage-specific sites obtained from those TFs (500-1000 sites) compared to CTCF (250 in total), although this may be explained by the greater degree of conservation in CTCF occupancy in general[269, 559], and between these two mouse subspecies in particular (see Chapter 2). This is in stark contrast to the results obtained in liver-specific transcription TFs using 6 replicates, where the effect was predominantly the result of *cis*-acting variation in the vast majority of cases (84-87%), whilst the effect from *cistrans*-acting variants was fairly marginal (More on that in section 3.3.6). The number of lineage-specific CTCF binding sites were equivalent between the two parental subspecies (120 and 130 for BL6 and CAST, respectively).

When non-lineage-specific *cistrans* CTCF binding variants where effects of *cis* and *cistrans* variants are found on both F0 and F1 alleles (3958) were classified to compensatory and diversifying modes of selections, these sites were distributed equally between the two categories (52% vs 48% respectively) (Figure 3.5d). Nonetheless, the differences between observed lineage- and non-lineage-specific CTCF binding sites in terms of their compensatory/diversifying modes of selections were not statistically significant. This suggests that CTCF sites influenced by *cistrans* effects on both alleles, and shared among both subspecies of mice, are neutral and do not particularly favour

either mode of selection, at least at the level we could detect with our available data. Taken together, these results indicate that novel CTCF binding sites are formed, either directly or indirectly, mostly via the contributions of variation in *cis*.

3.3.4 Dominant inheritance affect *cis*-directed CTCF occupancy

Previous work has shown inherited TF binding sites to be expressed in an additive or non-additive (dominant) fashion[257, 542, 594]. Additive inheritance was observed when the combined binding intensity of the two F1 alleles is equivalent to the sum of the two parental F0 alleles, whereas dominant (non-additive) inheritance when the total allelic binding signal from the F1 alleles is equal to that of either F0 parent. Dominant inheritance; therefore, could also be sub-categorized as "high" if the signal from the F1 alleles equals that of the parent with the higher binding intensity, or "low" if that signal is similar to the one from the parent with the lower binding intensity signal (Figure 3.6a). Inheritance in this context is defined by the total allelic signal from each replicate, whereas regulatory categories discussed above were assigned based on the ratio of signal between the F1 alleles and the ratios of their F0 parent of origin.

As with regulatory and lineage-specific category assignment, we fitted statistical models to test the three inheritance patterns outlined above, using Bayesian Information Criteria (BIC) to assess the outcomes in manner equivalent the one reported by Wong *et al.*[257] (see Methods). Of the CTCF binding sites under *cis*-influenced variants, 1021 passed the BIC difference minimum of 1 for inheritance pattern assignment, and 178 of the *trans*-acting variants were assigned an inheritance pattern with BIC > 1. In both cases of *cis*- and *trans*-acting variation, the predominant form was additive inheritance in which the total allelic signal from the F1 was equal to the sum of both parental allelic signals (55% and 43% respectively) (Figure 3.6b). The contributions of dominant inheritance of the inheritance observed in *cis*- and *trans*-influenced CTCF sites; however, were not equal. Most dominantly inherited *cis*-acting CTCF sites belonged to the dominant *high* variety of non-additive inheritance, in which the total allelic signal from the F1 was equal to that of the parent with the higher binding signal (34% vs 11% for dominant *low*). The same was observed, albeit to a much smaller scale, in *trans*-influenced CTCF sites (36% *high* vs 21% for *low*). When stratified by the F0 parent in with the higher median binding intensity ($F0_{MAX}$), the general distribution of inheritance modes did not differ between BL6 and CAST, and all trends in total, *cis* and *trans* were consistent in both mouse subspecies, and reflected the overall pattern (Figure 3.6b). A slight enrichment of sites inherited in dominant *low* form in BL6 in *trans* was observed, but owing to the small numbers involved, this might be a small number effect.

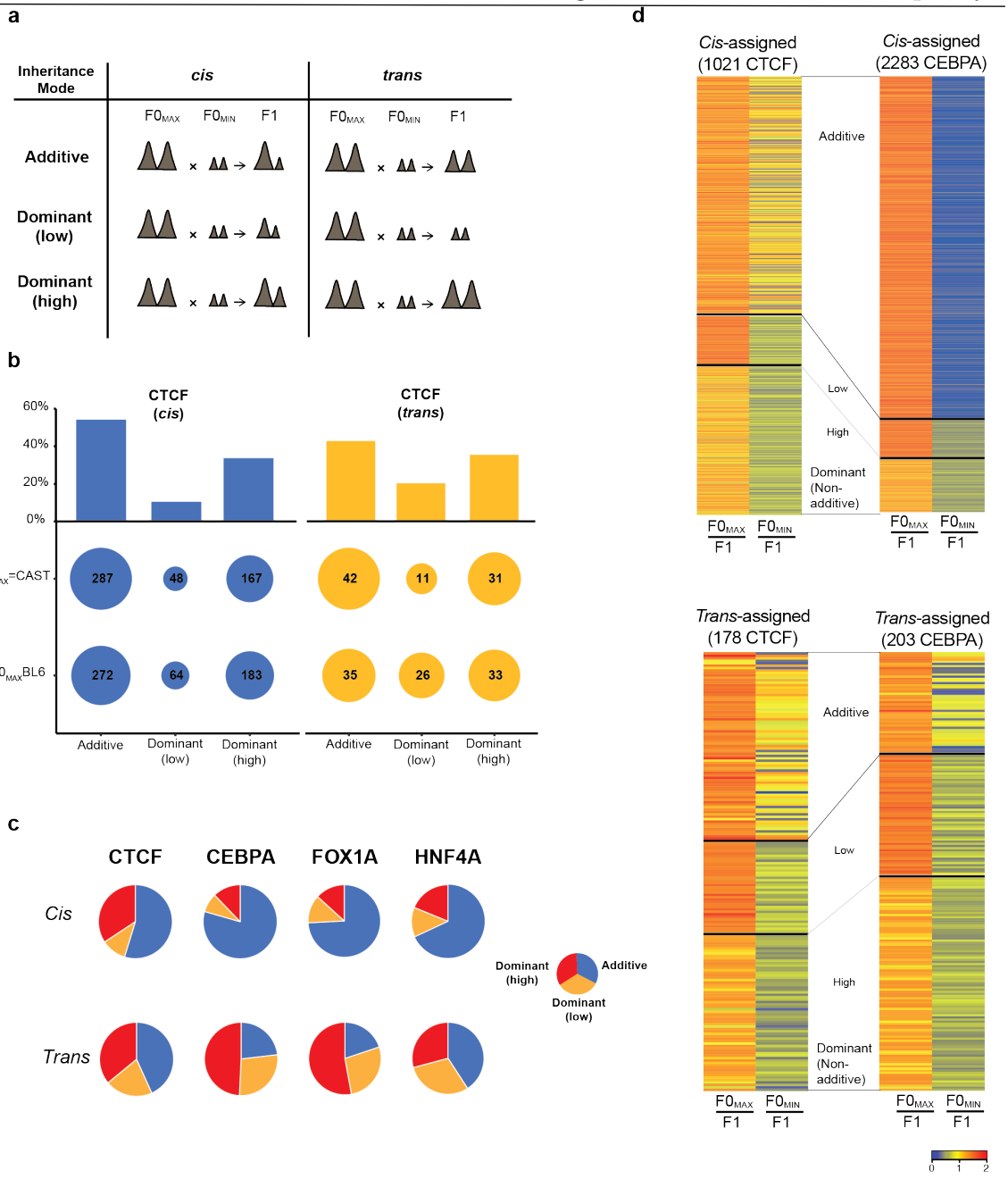


Figure 3.6: CTCF occupancy affected by *cis*-acting variation is show higher dominant effects

a A schematic model for assigning modes of inheritance for the *cis*- and *trans*-influenced TF binding sites. $F0_{MAX}$ and $F0_{MIN}$ refer to the F0 parental subspecies with the higher and lower median binding intensity, respectively. Binding intensities were summed across replicates in F0, and across alleles for F1. **b** The bar plots (*top*) show the proportion of *cis*- and *trans*-acting variants in CTCF binding sites based on their assigned mode of inheritance. The circle plot (*bottom*)

breaks down each mode of inheritance by the F0 father of origin with the higher median binding intensity ($F0_{MAX}$), with the number of sites per subspecies denoted inside the circles. The radius of each circle indicates the proportion of that mode of inheritance for the particular $F0_{MAX}$ parent for the same category of *cis*/*trans* variation. **c** Pie charts showing the relative proportions of the three modes of inheritance for CTCF and 3 liver-specific TFs as determined using a statistical model to fit binding sites affected by *cis* or *trans* variation (see Methods). **d** A heatmap showing CTCF (*left*) binding events affected by *cis*- and *trans*-acting variation. Different modes of inheritance were defined in **a** (see Methods for the statistical model). The data from CEBPA assigned modes of inheritance (*right*) for both *cis*- and *trans*-acting variation was used for comparison (see Methods). Total F1 counts were individually scaled to 1 (yellow).

This observation, in the case of *cis*-acting variation, is significantly different from the pattern observed in other TFs, where although the most prevalent mode of inheritance was additive (Figure 3.6c), the contribution of dominant inheritance was much reduced (χ^2 test for pairwise comparison between CTCF and other TFs with Bonferroni's correction, all p-values $<$ 2.2e-16). Although non-additive inheritance was the predominant form for *trans*-acting variation in CTCF and other TFs, the contributions of additive and both forms of non-additive inheritance varied in TF-specific fashion. For example, the enrichment of the dominant *high* mode of inheritance in *trans*-acting variation observed in CTCF was not seen in HNF4A, in which the dominant *low* mode was more common (Figure 3.6c).

A close inspection of the ratios of the signal in CTCF *cis* and *trans* in comparison with CEBPA reveals that the overall effect of regulatory variations in additive inheritance tends to centre the ratios of $F0_{MIN}$ to F1 towards 1, with few sites showing enrichments towards the lower ends (Figure 3.6d). Only dominant inheritance of the low variety shows a clear difference of the pattern in those ratios between $F0_{MAX}$ and $F0_{MIN}$ in *cis*-influenced inheritance (Figure 3.6d). This indicates that in CTCF even binding sites classified as influenced by *cis*-acting variation are under a clear dominant *low* influence that skews the inheritance pattern from the expected additive mode, which in turn could explain the increased effect of non-additive inheritance observed in the ratios of CTCF total binding signals compared to other TFs.

3.3.5 C*is*/*trans* CTCF binding is associated with higher occupancy conservation across tissues

Tissue-shared binding of CTCF is usually an indicator of both evolutionary conservation, and potential functional implication in regulatory activities. We thus

looked at the pattern of occupancy conservation of *cis/trans*-influenced variation on CTCF sites in previously described mouse ENCODE tissue CTCF libraries (see Chapter 2). 15.21% of all liver derived CTCF binding sites under *cis/trans* variation had their occupancy conserved in all 12 ENCODE tissues (over 2000 sites) (Figure 3.7a). A thousand more sites have shared occupancy in at least 11 tissues (7%). 5.44% of all *cis/trans* sites are bound in another tissue only, mostly in the kidney, whilst 8% have no observed binding in any other tissue but the liver. Out of a total of over 14000 *cis/trans* CTCF binding sites, 7 tissues exhibit shared occupancy of at least 5000 of these binding sites (Figure 3.7a). Similar results were obtained when the whole peak sequence of the SNV-containing binding sites were considered.

TF binding sites that are involved in tissue-wide regulatory functions are known to be under increased selective pressure, which manifest in the form of elevated levels of shared occupancy of their binding sites across tissues[399]. We evaluated the strength of each *cis/trans* CTCF binding sites in all of the 12 mouse ENCODE tissues, and used the Shannon Diversity Index[578], to characterise the diversity of *cis/trans* CTCF binding in terms of their abundance and conservation. *Cis/trans* CTCF binding conservation estimates derived from the analysis outlined above were additionally used to illustrate how changes in occupancy conservation across tissues track with the diversity of binding instance in varying numbers of cell types.

The high Shannon index value observed across tissue attest to the great degree in CTCF occupancy conservation, and is strongly correlated with *cis/trans* binding in the 12 ENCODE tissues included. The values, albeit generally high, form a bi-modal distribution, with the 5 tissues with fewer than 5000 *cis/trans* shared CTCF sites (seen in Figure 3.7a) forming a cluster towards the lower range of the Shannon index, and the top 7 tissues with greater tissue-sharedness in occupancy forming the cluster at the higher end of the distribution. These differences are mirrored in the degree of binding conservation of CTCF sites from 25% at one end to ~ 80% at the other end (Figure 3.7b).

The pattern of increased CTCF occupancy across tissues, however, does not hold true for the subset of *cis/trans*-influenced sites classified as lineage-specific (see section 3.3.3). Whereas a minority of the general pool of *cis/trans* CTCF sites (~8%) were only bound in the liver, 36% of lineage-specific sites were found to be liver-specific (Figure 3.7c). This is further reflected in overall lower number of shared sites across tissues. For example, of all *cis/trans* CTCF sites, 50% had shared binding in a minimum of 7 tissues. A similar fraction of sites was found to be bound in only three other tissues in the lineage-specific subset of cis/trans CTCF sites(Figure 3.7c).

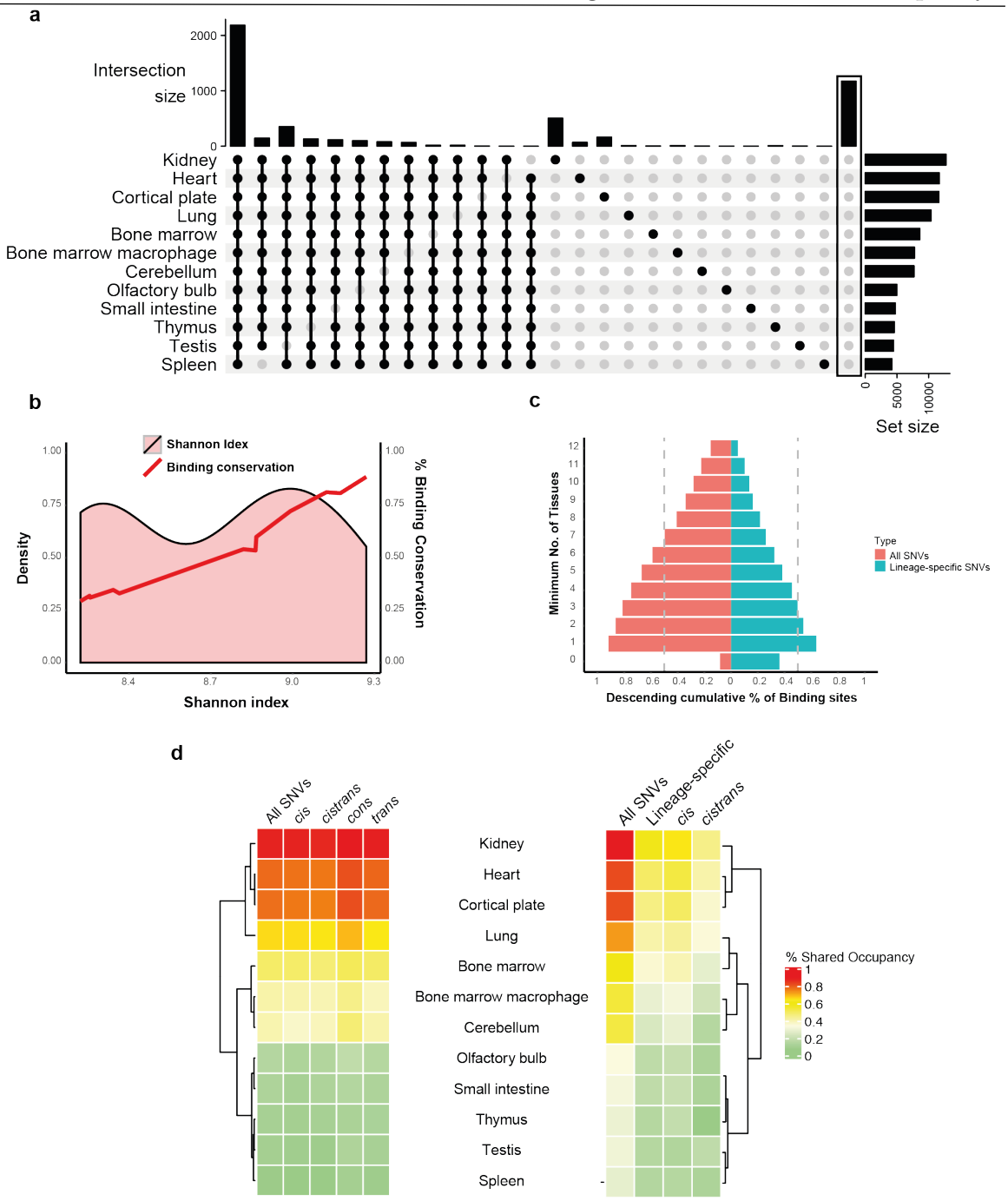


Figure 3.7: *Cis/trans* CTCF site exhibit higher binding conservation across all tissues

a UpSet plot illustrating the number of tissue-shared/specific *cis/trans* CTCF binding sites across the 12 mouse ENCODE tissues. The number of binding sites bound at each combination of tissues is indicated on the y-axis on the top bar chart. The original plot was reduced to only these 26 combinations to highlight

only highly tissue-shared and tissue-specific sites (original number of combinations with a minimum of 1 site was 613). The last bar (*right, boxed*) shows the number of CTCF sites that were tissue-specific to the liver. **b** Plot representing the association between *cis*/*trans* CTCF binding and occupancy conservation across 12 tissues. The density plot shows the frequency of CTCF shared binding in different tissues, whose diversity values are indicated on the x axis, calculated using the p-value estimates of *cis*/*trans* peak calls (see Methods). The higher the value on the x axis, the higher the number of CTCF sites bound. The red line represents the proportion of conserved CTCF binding within each bin of Shannon index. **c** Bar plot elucidating the fraction of *cis*/*trans* CTCF binding sites shared in increasing number of tissues for all SNV-influenced sites and lineage-specific ones. The x axis indicates the decreasing cumulative proportion of binding sites found at the minimum number of tissues on the y axis. The dashed grey line denotes the minimum number of tissues at which 50% sites are shared. **d** Heatmaps showing the proportion of tissue-sharedness in the different regulatory categories of CTCF binding (All on the left, lineage-specific on the right).

The patterns of tissue-wide conservation (in the case of general *cis*/*trans* sites) and tissue-specificity (in lineage-specific *cis*/*trans* sites) are mirrored across the underlying regulatory categories of which they are composed. All four *cis*/*trans* sites show the same level of occupancy conservation in the same tissues, with a slightly higher enrichment for the *cons* sites across the top ranking 7 tissues (Figure 3.7d *left*). No statistically significant difference between these categories were observed ($\chi 2$ test, p-value = 0.9091). Although lineage-specific sites show a general lack of occupancy conservation in other tissues, 41-59% of those exhibit shared occupancy in at least the top 4 tissues, kidney, heart, cortical plate and lung (Figure 3.7d *right*). The effect is strongest in lineage-specific *cis* CTCF sites, where tissue-shared occupancy ranges between 44-64% in the top 4 tissues (Figure 3.7d *right*). These differences, however, reflected the differing proportion of these two categories, and no statistically significant difference between them were observed ($\chi 2$ test, p-value = 0.8768).

3.3.6 The inclusion of biological replicates improves outcomes of analysis on *cis*/*trans* variation in TFs

As the results from the subsampling strategies, explored in section 3.3.1 earlier, had indicated, adding extra biological replicates to the analysis prove useful in enhancing our ability to call the *cis*/*trans* regulatory region in TF binding sites more confidently and significantly improve our estimation of the true proportion of these categories. This was particularly apparent in the case of TF binding sites under influence from

cistrans variants, whereby the availability of more ChIP-seq libraries helped resolve them into their true *cis*/*trans*/*cons* categories (Figure 3.3).

As a concluding section for the work done in this chapter, an investigation was carried out into the different facets the addition of more replicates improves in our analysis. By randomly selecting 2, 3, 4 and 5 replicates out of the originally available 6 for all three TFs, CEBPA, FOXA1 and HNF4A, we looked at how these changes reflect on the various aspects originally studied. As seen previously, the progressive addition of replicates does change the proportion of the *cis*/*trans* regulatory classes of TF ChIP-seq signals (Figure 3.8a). The *cistrans* category is generally reduced in proportion with increasing number of replicates, although this change is not completely uniform. At 2 replicate, they make up the second most common type of regulatory variants, but starting from 3 replicates onwards, they are progressively reduced in numbers and the conserved (*cons*) variants take their place as the second most common type of variant in FOXA1 and HNF4A. This; however, was not the case for CEBPA, where they remained the 2nd most common type at 3 and 5 replicates. This, nonetheless, maybe a stochastic effect arising from the random selection of libraries for this analysis, and different set of libraries at 3 and 5 replicates may reflect the general pattern seen in the other two TFs. The reduction in the number of *cistrans* sites invariably lead to the enrichment of the three other categories, particularly the *cons* and *trans* sites. The changes in numbers estimated from adding replicates in all four *cis*/*trans* categories was consistently statistically significant ($\chi 2$ test, all p-values $<$ 2.2e-16).

These changes were not limited to the overall number of sites within each category alone, but also extended to other features of our ability to confidently assign variants to their appropriate category. As explained in the Methods section of this chapter, category assignment was carried out using Bayesian Information Criteria (BIC), by looking at the difference in value from the category with the lowest BIC value to the second lowest. This differential BIC value can be used as a measure of the confident of variant category assignment; the higher the value of BIC, the more reliable the call.

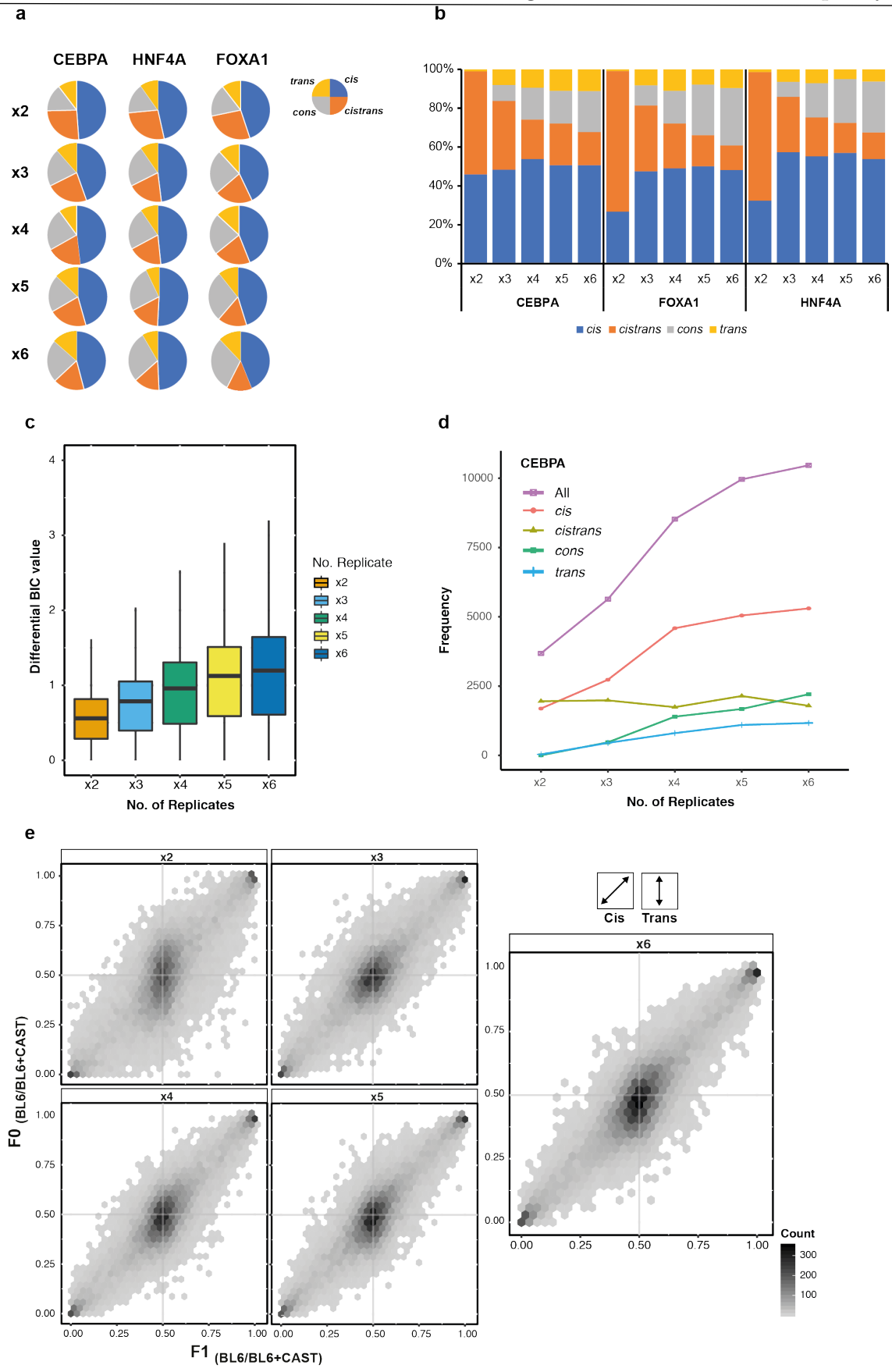

a
CEBPA
HNF4A
FOXA1
x2
x3
x4
x5
x6
trans
cis
cons
cistrans
b
100%
80%
60%
40%
20%
0%
x2 x3 x4 x5 x6 x2 x3 x4 x5 x6 x2 x3 x4 x5 x6
CEBPA
FOXA1
HNF4A
cis cistrans cons trans
c
Differential BIC value
No. Replicate
x2
x3
x4
x5
x6
No. of Replicates
d
CEBPA
All
cis
cistrans
cons
trans
Frequency
10000
7500
5000
2500
0
No. of Replicates
e
x2
x3
x4
x5
x6
F0 (BL6/BL6+CAST)
F1 (BL6/BL6+CAST)
Cis
Trans
Count
300
200
100

Figure 3.8: The addition of extra biological replicates enhances category assignment and BIC estimation

a Pie charts displaying the regulatory variant make-up of 3 TF binding sites under *cis*/*trans* influence derived from randomly selecting an increasing number of biological replicates. **b** 100% stacked bar chart of the fraction of each *cis*/*trans* category that are called with a minimum BIC of >=1 in 3 TFs in increasing number of replicates. **c** Boxplots of the differential BIC values (see Methods section 3.2.2.8) for all CEBPA binding sites under *cis*/*trans* regulatory variation in ascending number of libraries. **d** Line plots of the number of *cis*/*trans* TF binding sites called at BIC >=1. The "All" category represents the total number of sites from all four *cis*/*trans* categories present at each number of replicates. The plot shows the data for CEBPA**. e** Hexagonal heatmaps for the mean values of F0 versus. F1 binding intensity ratios (BL6 vs. CAST) for every *cis*/*trans* CEBPA site in 2 to 6 biological replicates.

As the results from re-running the analysis with ascending number of replicates have shown, the availability of extra libraries does increase the overall number of TF sites whose assigned BIC value is >=1, and provides a more balanced representation of the four categories at variants called with higher confidence (Figure 3.8b). At 2 replicates only, only a subset of almost exclusively *cis* and *cistrans* variants have BIC values >=1, with *cistrans* being the most abundant sites with higher confidence calls. The situation immediately improves at 3 replicates, with *cons* and *trans* sites now being present, albeit at frequencies that do not reflect their overall proportion of the total set of TF binding sites. This increase from 3 replicates onwards, however, invariably happens at the expense of the *cistrans* variants, whose fraction at BIC >=1 is progressively diminished (Figure 3.8b). The increase in the overall number of TF binding sites called with higher confidence is not limited to sites with a BIC value >= 1. The overall BIC value estimation of the all sites improves markedly with ascending number of biological replicates (Figure 3.8c). The changes in the BIC estimations of all TF binding sites were found to be strongly statistically significant (Kruskal-Wallis test, p-value < 2.2e-16). The same results were observed for the other two liver-specific TFs (see Appendix 2, Figure S2.1).

The changes seen in differing proportion of *cis*/*trans* category at higher BIC values with increasing replicate number could be the result of either more variants being called with higher reliability whilst the number of high-BIC *cistrans* sites remains roughly the same, or due to the resolution of some these sites into other categories. To address this question, we looked at the frequency of all sites with a BIC >= 1 at each number of replicates (Figure 3.8d). As the results indicate, it does look like the former explanation is the one the data supports. As more libraries are added, more sites have

their variants assigned at higher BIC value in general and particularly in *cis*, *cons* and *trans* ($\chi 2$ test, p-value < 2.2e-16). The number of *cistrans* variants with BIC >=1 remains effectively the same (Figure 3.8d). This indicates that even though the overall results of *cis*/*trans* assignment generally improve with the addition of more input data, *cistrans* calls made with higher confidence are not affected, and do not resolve to variant calls to other categories.

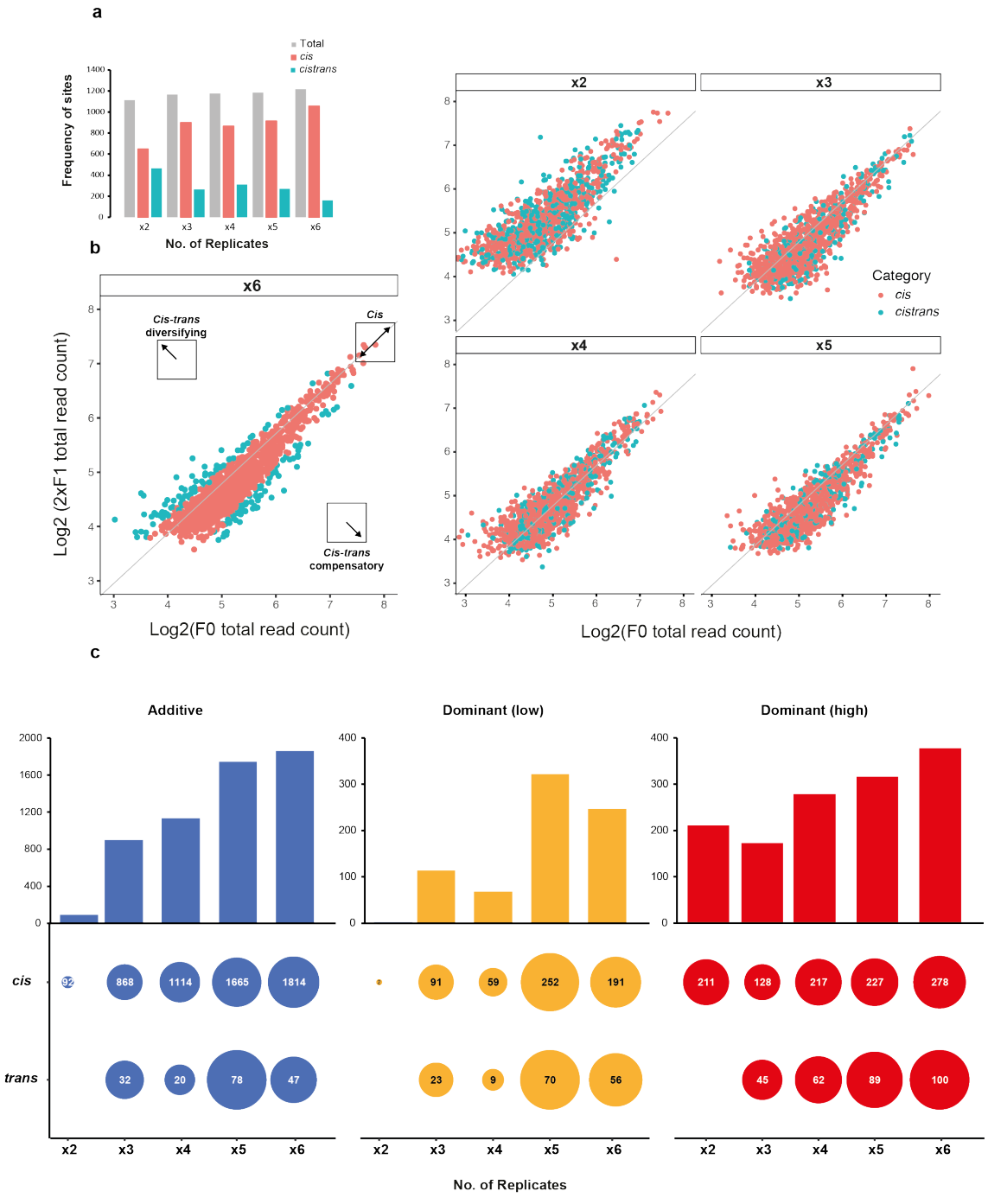


Figure 3.9: Availability of more libraries markedly improves estimates for lineage- specificity and inheritance patterns of in *cis*/*trans* TF sites

a Bar chart showing the number of TF binding sites classed as lineage-specific for both *cis* and *cistrans* variants in 2 to 6 biological replicates. The "Total" bar is equal to the sum of the two regulatory categories. The plot shows the data for CEBPA**. b** Scatter plots of average CEBPA log2 F0 total read counts against average log2 F1 read count (BL6 plus CAST allele) multiplied by 2, using averages in 2 to 6 biological replicates. TF binding sites affected by *cis*-acting variants are expected to correlate between the F0 and F1 along the diagonal (*grey line*). TF binding sites affected by *cistrans*-acting variants disperse further away than the expected straight line, reflecting their direction of lineage-specific variation. **c** Bar plot (*top*) of the number of CEBPA binding sites based on their assigned mode of inheritance in ascending number of biological replicates. Note the different y-axis for each between the "Additive" and "Dominant" inheritance modes. The circles (*bottom*) illustrate the make-up of each mode of inheritance by the type of *cis*/*trans* variation acting on the binding site for each number of replicates, with the number of sites per category denoted inside the circles. The radius of the circles encodes the relative number of sites compared to the numbers observed across different number of replicates for the same category of *cis*/*trans* variation. No *trans* sites passed criteria for inclusion in this analysis (See method) for 2 biological replicates.

As with the differences in category assignment and higher confidence calls, the availability of extra biological replicate improves other aspects of the analysis explored in this chapter. This is, for example, evidenced in the distribution of *cis*/*trans* TF binding sites differences between the two subspecies (Figure 3.8e). Although the overall pattern observed between the ratios of F0 and F1 read enrichments is strongly correlated, these ratios exhibit a higher degree of dispersion and deviation from the linearity at 2 replicates for CEBPA, but it clearly improves with the addition of more libraries. This improvement is reflected in higher R values for Pearson correlation coefficients from 2 to 6 replicates (R = 0.85, 0.9, 0.91, 0.92, 0.93 respectively, all p-values $<$ 2.2e-16). It is worth noting that even at 2 replicates only, the R correlation coefficient for CEBPA was still statistically significantly different from that of CTCF (R = 0.7, t-test, p-value $<$ 2.2e-16). Results from the other two TFs produced the same pattern (see Appendix 2, Figure S2.2).

The inclusion of more biological libraries additionally has a defining effect on the ability to distinguish the regulatory variation driving the evolution of lineage-specific TF binding sites. As defined in 3.3.3, lineage-specific binding sites are a subset of sites that are bound exclusively in subspecies-specific manner in one F0 parent, and whose

allelic read enrichment in F1 hybrids is specific to the parent of origin allele. By virtue of this definition, only *cis* and *cistrans* variants can be classed as lineage-specific (see section 3.3.1). When this definition was applied to the binding sites of CEBPA in a series of analyses using 2 to 6 replicates, we obtained two key findings. Firstly, the total number of sites classed as lineage-specific does not apparently change much with the addition of extra data (Figure 3.9a). There was only 8% increase in the number of lineage-specific sites from 2 replicates (1113) to 6 replicates (1217). However, the biggest difference appears to be the regulatory variation acting on these sites. The number of lineage-specific *cis* sites gradually increase with every additional library until it reaches 87% of all lineage-specific sites. This rise coincide with the reduction in *cistrans* lineage-specific sites from 40% (in 2 replicates) to only 13% (16 replicates) (Figure 3.9a). This change is statistically significant ($\chi 2$ test, p-value $<$ 2.2e-16). These changes did not show any preference towards a particular F0 subspecies and distributed roughly equally between the two parental subspecies across all replicate numbers.

The second notable outcome of looking at lineage-specific binding in increasing number of replicates is that even though the total number of sites did not seem to change, they still display a different pattern when their parental read enrichments were visualised against their hybrid offspring (Figure 3.9b). At 2 replicates, there is no clear distinction between *cis* and *cistrans* CEBPA binding sites and the read enrichment values between parents and offspring is skewed further away from the linear fashion seen in 6 replicates. This quickly improves with the addition of an extra replicate, and from there on, the distribution takes a more uniform shape. The distinction between *cis* and *cistrans* values, however, remains generally indistinguishable and only begin to resolve at a much higher number of replicates (5) (Figure 3.9b). The same pattern was observed in the other two TFs, although the number of lineage-specific sites varied in a TF-specific manner (see Appendix 2, Figure S2.3).

The effect of incorporating additional biological replicates can also be seen in teasing apart the inheritance pattern of *cis*/*trans*-influenced TF binding sites. Similarly to CTCF, the distribution of inheritance patterns of CEBPA binding sites between F1 and F0 in 2 replicates shows signs of increased dominance (*high*) in its occupancy pattern that corresponds to stronger F1 occupancy levels compared to parental measurements (Figure 3.9c). Even though additive inheritance starts to become the predominant mode from 3 biological replicates onward for *cis* variants, contributions from the dominant (*low*) only reach their relative proportion in higher number of replicates (>4). Notably, roughly the same number of *cis* TF sites are inherited in the dominant *high* mode regardless of the replicate number (Figure 3.9c). The pattern of *trans* sites inheritance in ascending number of replicates appeared to vary considerably.

This might be due to their generally smaller number combined with the stochastic effect of the random subsampling of replicates. A clear pattern of increased number of *trans* sites that met the threshold to for inheritance mode assignment is apparent with the addition of extra biological replicate. The same results were observed for the other two liver-specific TFs (see Appendix 2, Figure S2.4). This suggests that to tease apart the true contribution of the different modes on inheritance in *trans*-acting variation, a higher number of biological replicates is required.

3.4 Discussion

Regulation of gene transcription is the outcome of complex interplay between TFs and the genes they regulate. This regulation is at least partially dependent, and explained, by variation in genome sequence affecting TF occupancy and gene expression. Regulation of transcription can be mediated via *cis*-acting elements (e.g. promoters, enhancers, TF binding sites, etc.) or *trans*-acting elements (TF, diffusible elements, ncRNAs, nuclear environment, etc.)[614]. The cross-talk between *cis*- and *trans*-acting variation generates transcriptional regulatory circuits that process this complex array of information, and produce robust gene expression[615]. Whereas most TFs show a high degree of variability in their occupancy among genetically identical individuals[616], and an accelerated evolutionary divergence of their bindings[394], CTCF is notable for exhibiting a high level of binding conservation[269], particularly around genomic features of direct transcriptional involvement[399]. Previous work has focused on how genetic sequence variants correspond to TF binding differences between alleles in mouse and human cell lines[257, 260, 601, 615, 617]. However, to out knowledge, CTCF occupancy difference in response to genetic sequence variation in *cis*/*trans* has not been reported.

In this study, we have interrogated the mechanisms underlying sequence variation effect on CTCF occupancy using a hybrid mouse model. This model had previously been used to examine *cis*- and *trans*-acting variation influence on TF binding and gene expression, where differences in TF occupancy were mainly the product of variation acting in *cis*[257].

Due to the disparity in biological replicate availability between our study and the one in Wong et al.[257], we re-ran comparative analyses between CTCF and three liver-specific TFs (CEBPA, FOXA1 and HNF4A) on two randomly selected biological replicates from the 6 replicates report previously for these TFs to allow for meaningful comparisons. Our results demonstrably indicate that quantitative differences in CTCF binding sites are driven by a significant contribution by *trans*-acting elements.

Whereas the effect of genetic differences acting in *cis* and *cons* contributed to 63-64% of differences in occupancy of liver-specific TFs, the effect of these variants is only 55% in CTCF. The remainder of effect comes from either *trans*- or *cistrans* variation acting on CTCF binding sites. However, the results obtained from 2 replicates in liver-specific TFs show an increase in *cistrans* assignment of regulatory variants when compared to the original study. Our subsampling approach indicated that this had come at the expense of both the *conserved* and *cis* variants. However, the results of subsampling support our observations of fewer *cis* and more *trans* regulatory variants in CTCF as consistently different from the proportions observed in liver-specific TFs in ascending number of replicates.

Although CTCF binding has been shown to exhibit higher degree of conservation than most TFs[269, 559, 618], and particularly between these two mouse subspecies (See Chapter 2), this conservation of binding does not necessarily translate to either conservation of sequence (number of binding sites with SNVs), or signal (a reduction in the sites classified as conserved/*cis* compared to liver-specific TFs). The contribution of *trans*-acting variants has a substantial effect on individual allelic signal in F1. Notably, CTCF occupancy transmission from parent to offspring in mouse was equivalent to the rate observed in human lymphoblastoid cell line (LCL) ($r = 0.66$, p-value = 6.3e-10) [152].

Comparison of various facets of *cis*- and *trans*-acting variation on CTCF occupancy, as compared to other TFs, has produced a number of fascinating insights. Allelic differences in CTCF occupancy at *cis*-affected sites do not correlate with the binding signal of neighbouring sites either on the long or short range. This in contrast to previous studies looking at TF binding coordination of binding in response to *cis*/*trans* variation in mouse, and analysis of local and distant correlation of gene expression and human eQTLs[257, 603, 613, 619]. The way the analysis was conducted, on the other hand, may have missed the long-range interactions driven by CTCF enrichment at chromatin contacts as these were not specifically selected for during this analysis.

Analysis of *cis* and *cistrans* variation on lineage-specific binding of CTCF exhibited a similar pattern to that seen in liver-specific TFs. CTCF occupancy; however, showed a substantial difference in terms of the number of lineage-specific sites (250 in total) compared to those obtained from those TFs (500-1000 sites). This is likely a consequence of the greater degree of conservation in CTCF occupancy in general[269, 559], and between these two mouse subspecies in particular (see Chapter 2). Previous work demonstrated that new binding sites rise and become fixed on microevolutionary timescales under assumption of neutral evolution, and both compensatory and diversifying *trans* effects should be equally favoured[541, 620]. This

was further confirmed with previous work on liver TFs whose subspecies-specific binding was influenced equally by both compensatory and diversifying modes of *trans*-acting influence[257].

The effect of increased *trans*-acting variation on CTCF is further seen when the pattern of allelic signal inheritance was investigated. The predominant form was additive in both cases of *cis*- and *trans*-acting variation. Nevertheless, most dominantly inherited *cis/trans*-acting CTCF sites came in the form of dominant *high*, in which the CTCF occupancy in hybrid offspring corresponded to that of the parent with the higher binding intensity. This was, particularly in the case of *cis*-acting variation, notably different from the mode of inheritance observed in liver-specific TFs, in which even though the additive mode of inheritance was most prevalent, the contribution of dominant inheritance was significantly reduced. We additionally observed that even CTCF binding sites influenced by *cis*-acting variation are under a clear dominant *high* influence that skews the inheritance pattern from the expected additive mode, suggesting that it could be driving the increased effect of non-additive inheritance observed in CTCF compared to liver-specific TFs. This can be explained by the pervasiveness of *trans*-acting variants at the higher and lower ends of the distribution of binding signal values driving the ratios between the total signal of F1 alleles to either parent towards the middle, diluting the effect from *cis*-acting variants. The differences in inheritance pattern also extend to *trans*-influenced CTCF binding sites, where even though these sites are dominantly inherited, their dominant inheritance comes in the "high" variety. Indeed, a study by Stergachis *et al.* proposed that selection on gene regulation during mammalian evolution is specifically targeted at the *trans*-regulatory network level, enabling potential *cis*-regulatory plasticity[618].

As we discussed in Chapter 2, conserved CTCF binding exhibits higher tissue-sharedness in its occupancy than evolutionary young, subspecies-specific sites. We investigated the extent of occupancy conservation across cell-types in CTCF sites subject to *cis/trans* regulatory variation. Similar to *musculus*-common sites, *cis/trans* CTCF sites, regardless of their regulatory category, exhibit a substantial degree of binding conservation across tissues. The type and number of tissues in which these sites bind is similar to the pattern observed in conserved CTCF sites. Even though lineage-specific *cis* and *cistrans* are restricted in their tissue distribution, some are still bound across multiple tissues. These findings support the findings from Chapter 2, and further illustrate that conserved sites, even when influenced by sequence variants in the binding site, remain strongly associated with tissue-wide binding. Lineage/subspecies-specific sites, on the other hand, are more restricted in their binding across tissues, and only a subset with higher regulatory potential manage to exhibit tissue-wide occupancy.

The findings of this investigation; nonetheless, do have their limitations. First, only two biological replicates were obtained for each of the reciprocal crosses to produce hybrid F1 mice. More replicates have the potential to either enhance or reduce the effect sizes we measured in this analysis. We performed an exploratory analysis into the effect of inclusion of extra biological replicates on the outcomes of some of the analyses carried out in this chapter. By randomly selecting 2, 3, 4 and 5 replicates out of the originally available 6 for all three TFs, CEBPA, FOXA1 and HNF4A, we observed changes in the proportion of the *cis*/*trans* regulatory classes of TF. The proportion of TF sites assigned to the *cistrans* category decreased gradually with increasing number of replicates. This reduction was most likely due to the resolution of low-confidence *cistrans* sites into the three other categories.

The re-assignment of low-confidence *cistrans* sites was further evidenced by analysis of the number and proportions of TF *cis*/*trans* sites whose category-assignment calls were made with higher-confidence. We showed that the incorporation of more biological replicates not only enhanced our ability to assign categories more reliably (particularly conserved and *trans* sites), the category assignment confidence scores significantly improved with increasing replicate number. Whereas the total number of high-confidence conserved, *cis* and *trans* TF sites increased with increasing replicate number, high confidence *cistrans* sites remained roughly the same. This indicates that the general reduction in *cistrans* sites in higher number of replicates does not affect those sites, but it is the lower quality category-assignment that resolve into the other categories when more data is made available for the analysis.

Furthermore, we observed an improvement in *cis* effect sizes in the ratios of TF binding signal in hybrid offspring compared to the parental subspecies when extra replicates are integrated into the analysis. Again, this is most likely the result of the increase in sites assigned as conserved and/or *cis* due to the resolution of *cistrans* sites. Similarly, although the number of lineage-specific *cis*/*trans* TF binding sites do not change with increasing replicate number, the ratios of *cis* to *cistrans* do, again due to the re-assignment of sites previously categorised as *cistrans*. Additionally, the inclusion of more biological libraries has a marked effect on our ability to discern the regulatory variation underlying the evolution of lineage-specific TF binding sites.

We also observed a clear effect on teasing apart the inheritance pattern of *cis*/*trans*-influenced TF binding sites when incorporating additional biological replicates. Similar to CTCF, even though additive inheritance remains the predominant mode of inheritance for *cis* variants in 2-6 replicates, liver-specific TFs shows signs of a dominant effect (*high*) in their occupancy pattern. However, contributions from the dominant (*low*) form of non-additive inheritance only manifest the proportion reported in Wong et al.[257] in higher number of replicates. This

indicates that establishing the correct modes on inheritance in *cis*- and *trans*-acting variation, a higher number of biological replicates is generally required.

It is worth mentioning that most of the CTCF sites, similar to other TFs, had no informative SNVs that can be used to discern differential allelic signal in the F1, so the full extent of this effect is not fully resolved. The prevalence of CTCF binding sites affected by variation acting in a *cis*/*trans* fashion complicates the pattern of impact on differential allelic signal in the F1. Adding more replicates or deeper sequencing may allow for the partial resolution of this band of regulatory elements, or alternatively emphasize their effect on the binding of CTCF. Lastly, all definitions of regulatory classes of sequence variations effects on CTCF and other TF bindings were based on statistical models. These sites in their nuclear environment are naturally biologically heterogeneous, and the spectrum of effects on their occupancy is potentially wider.

Nevertheless, this study provides further support to the use of this model to investigate the differential binding of TFs in hybrid nuclear environments to discern the various effects that influences their occupancy and the selective forces at work. This work, along with previous studies[152, 257, 260], shows the value of this approach in studying the roles genetic variation plays in TF binding regulation of gene expression.

Chapter 4

Regulatory potential of CTCF binding in closely-related mice

4.1 Introduction

Repetitive elements have played a major role in the in shaping the regulatory sequence of the non-coding genome throughout the evolution of mammalian lineages through the creation of novel loci for the binding of transcription factors[387, 560-562]. CTCF is a prime example of this mechanism, where waves of short nuclear interspersed elements (SINEs) expansions spread the CTCF binding site within the mouse lineage[559, 621].

CTCF binding, coupled with interaction with the cohesin complex of proteins, is a major component of the process that establishes and maintains the integrity of the 3D genome structure[622, 623]. CTCF binding colocalization with the formation of the cohesin complex is part of the process of chromatin loop formation through a loop extrusion mechanism, whereby the loop anchors are marked with two bound CTCF molecules that help stabilise the loop, resulting in the establishment of topologically associating domains (TADs)[227, 235, 285, 577, 624, 625]. TADs have reportedly been highly conserved during mammalian evolution, appearing invariably at consistent genomic loci across species and cell-types[235, 312, 626].

The presence of TE-derived CTCF binding sites in evolutionary young sites, along with CTCF association with TADs in chromatin loop anchors provide the basis for possible evolutionary mechanism to forming novel higher order chromatin structures. The insertion of subspecies-specific binding sites by action of SINEs in the

vicinity og more conserved CTCF binding sites gives opportunity for novel loop contacts to form. This model is supported by findings reporting TE-derived CTCF occupancy observed in association with chromatin loop anchors in mouse[219, 238].

CTCF-mediated chromatin looping is reported to allow distal enhancers, lying further away up/downstream of their target genes, to interact in physical space with target promoters to activate gene expression[355, 627, 628]. Cohesin-loading near promoters interacts with CTCF bound near distal enhancers to mediate promoter-enhancer contact. A study in ES cells demonstrated the potential of CTCF when interacting with cohesin complex near active regulatory elements to readily recruit the core promoter factor TAF3, and via the establishment of TAF3- dependent loop, CTCF facilitates promoter-enhancer contact[317].

In Chapter 2, we observed the expansion of these SINE-derived CTCF binding sites in two main clusters: an older expansion shared among sites in the *musculus*-common set of sites, and a recent one that took place after the divergence of both subspecies, and is found primarily evolutionary young sites. The results from Chapter 3 illustrated the contribution of *cis*- and *trans*-acting variation on the binding of CTCF in a set of sites shared between the two closely-related subspecies and characterised by informative SNVs. In this chapter, we present our analysis of at the regulatory and functional potential of CTCF occupancy, comparing and contrasting between CTCF sites differing in their evolutionary/tissue-specificity (from Chapter 2) and sites that exhibit binding variation in the form of *cis*/*trans* regulatory variants (from Chapter 3). Our findings provide evidence of dynamic evolutionary conservation in TAD-boundary association where conservation of binding coincides with substantial contribution of both sites under *cis*/*trans* regulatory variation and subspecies-specific CTCF sites. We also observed strong association between CTCF and cohesin-complex proteins in *cis*/*trans*-influenced sites and evolutionary young, tissue-shared sites equivalent to the level seen in more conserved sites. The regulatory and functional aspects of CTCF binding in all sets of binding sites considered in this thesis appeared to maintain the stability of pre-existing higher-order chromatin structures, whilst also providing a template for subspecies- and tissue-specific genomic innovation.

This investigation is the result of a collaboration between Dr. Paul Flicek's research group at the EMBL European Bioinformatics Institute and Dr. Duncan Odom's laboratory at the Cancer Research UK Cambridge Institute. I carried out the computational analysis, except where otherwise specified.

4.2 Methods

4.2.1 Repeat content in *cis/trans*-influenced CTCF sites

The full set of transposable elements (TEs) for the C57BL/6J mouse genome was retrieved from Thybert et al[559] and used to analyse repeat content in *cis/trans*-influenced CTCF and three liver-specific TFs (CEBPA, FOXA1 and HNF4A) binding sites. The contribution of repeat elements to the binding sites was estimated using the intersection between the TF (CTCF, CEBPA, FOXA1 and HNF4A) binding sites and the full set of the four TE superfamilies to calculate the fraction of sequence occupied in each binding site. We additionally used the full set of mouse TEs to calculate the background representation of the 4 most common superfamilies of TEs (SINEs, LINEs, LTRs, DNA transposons) in the mouse genome to compare to all four TFs.

A non-*cis/trans* TF binding sites set was derived for CTCF and the other 3 TFs to use for comparison of repeat content. A union peak file was generated for all ChIP-seq identified binding sites for each TF. BEDTools version 2.2.5.0[511, 512]with the option -v was used to filter out all binding sites shared with the SNV-containing, *cis/trans*-influenced set of binding sites.

BEDTools intersect 2.2.5.0 with the option -wo to return the length of sequence overlap between the peak sequence and the repeat, followed by division by the total length of the binding site to return the percentage of sequence occupied by repeat elements. To estimate the relative age of the repeat element in which a TF binding site is embedded, we used the percentage of sequence substitutions in each repeat from the consensus in the same way described in Chapter 2 Methods section 2.2.2.3. A random set of non-overlapping, chromosome-matched genomic sequences to *cis/trans* CTCF sites was generated for comparative analysis using BEDTools version 2.2.5.0 shuffle tool to generate sequences equal in number and length to the total number of CTCF (*cis/trans* and non-*cis/trans*) sites.

4.2.2 Gene feature analysis of CTCF sites

We utilised basic gene features derived from the most recent mouse genome assembly (GENECODE Release M23 (GRCm38.p6)[629]) to investigate the genomic distribution of CTCF binding sites classes, both on the basis evolutionary/tissue-specificity (Chapter 2) and *cis/trans*-acting variation (Chapter 3), in addition to the 3 liver-specific TFs mentioned above. We defined five main classes of features: Intergenic, Promoters, TSSs (transcription start site), Exons and Introns. We performed the genomic characterisation of TF binding sites using the annotatePeaks.pl

tool from HOMER (Hypergeometric Optimization of Motif EnRichment) suite (v4.11)[630].

CTCF binding sites regions were analysed with GREAT version 3.0[515] using default parameters to determine the distance from each CTCF site of each category to the nearest transcription start site (TSS). All CTCF sites more than ±100 kb from the nearest TSS were pooled together. CTCF proximity to downstream gene bodies was measured using PeakAnalyzer version 1.4[631], with annotation from the most recent mouse genome assembly (GRCm38).

4.2.3 CTCF occupancy at proximal active regulatory elements

Liver ChIP-seq libraries for H3K4me3 (a histone modification predictive of active promoter regions) and H3K27ac (a histone modification predictive of active promoters and enhancers[209]) were obtained for C57BL/6J (BL6) from Wong et al 2017[257], each with three biological replicates. Reads were aligned, filtered and peaks were called using the methodology explained in detail in Chapter 3 Methods section 3.2.2.1. Only peaks common in a minimum of two replicates were used to define active regulatory elements. A promoter region was defined by the localisation of either H3K4me3 only, or with overlapping H3K27ac signal, whereas enhancers were defined by the presence of the histone modification H3K27ac alone within the peak region.

Co-localisation of TF binding sites in regulatory elements was defined using an intersection of at least 1 bp between the TF binding site and the regulatory element. TF. BEDTools intersect 2.2.5.0 with the option -wa -wb to retrieve all overlaps between binding sites and active regulatory regions. The analysis was performed similarly for both TFs under *cis*/*trans* regulatory variation, and CTCF in different evolutionary/tissue-specificity classes.

CTCF proximity to active regulatory regions was measured using BEDTools closest 2.2.5.0, with the options -D ref and -mdb all against all active regulatory region to return only the closest enhancer/promoter but not both at the same time. We excluded any sites whose distance to the CTCF binding site is 0 (i.e. overlaps the binding site) as these sites have already been considered for the co-localisation analysis outlined earlier.

4.2.4 CTCF occupancy at TAD-boundary analysis

To calculate the distance from each TF binding sites to the nearest up/downstream topologically-associated domain (TAD) boundary, mouse liver TAD

boundary data from Vietri Rudan et al. 2015[238] were used. We defined the TAD boundary as the start or end nucleotides for every TAD in that dataset. The distance from each TF binding site and its nearest TAD boundary was measured using the BEDTools closest 2.2.5.0 tool. This was done for binding sites under *cis*/*trans* regulatory variation (CTCF and other TFs) and evolutionary common/young (CTCF).

Enrichment of TF binding sites within single TADs was measured by counting the number of each TF within TADs using BEDTools intersect 2.2.5.0 with the option -c. These resultant counts were compared against expected counts. As TADs vary widely in size, a unique expected frequency needed to be calculated separately for each TF. This TF-specific expected frequency was derived by dividing the total sequence length occupied by all TADs by the number of SNVs of each TF, producing the frequency in which a SNV is found in the genome. By dividing the length of each TAD by the particular frequency for each TF, an estimate of the number of SNVs/binding sites expected for each TAD individually.

4.2.5 CTCF recruitment of cohesin-complex proteins

ChIP-seq data for three cohesin-complex subunits (Rad21, STAG1 and STAG2) in liver, two biological replicates for each subunit, from adult male mice and matched controls were retrieved for BL6 from Faure et al. 2012[362]. Sequence reads were aligned to the GRCm38 reference genomes using BWA version 0.7.12[494] for each biological replicate and control. Cohesin-complex subunits regions were identified by peak calling from aligned sequence reads using MACS version 2.1.0[632] callpeak function with a p-value threshold of 0.001 and default parameters to call peaks representing cohesin-bound regions in the genome.

Genomic regions where at least two cohesin subunits peaks overlap were merged using BEDOPS version 2.4.3[633], and cohesin merged regions overlapping with *musculus*-common/BL6-specific/BL6-tissue-shared from our CTCF liver binding sites were identified. The intersection analysis was done for CTCF co-occupancy with two and three subunits, owing to the significantly fewer number of ChIP-seq peaks retrieved from the STAG1 data. The set of regions where a minimum of 2 cohesin subunits were bound was used for all further analysis involving CTCF (both *cis*/*trans*-influenced and evolutionary common/specific) and the liver-specific TFs.

When comparing cohesin-complex recruitment between CTCF and the liver-specific TFs, we considered two types of cohesin-bound regions. Cohesin-and CTCF (CAC) sites are defined as sequences where a CTCF site (*cis*/*trans* or otherwise) co-localises with a minimum of two overlapping two cohesin subunits. Cohesin-non-CTCF

(CNC) sites are sequences where the recruitment of two overlapping cohesin subunits was not associated with the binding of any CTCF.

To investigate the correlation between the evolutionary/tissue-specificity type and cohesin-recruitment by CTCF, we divided each set of CTCF binding sites into ten 10% bins based on descending ChIP-seq signal. ChIP signal in this context referred to the reads pileup per peak from the replicate where the peak signal was at its highest, for each of the three evolutionary/tissue-specific classifications: *musculus*-common/BL6-specific/BL6-tissue-shared. Signal intensity was then compared to the level of cohesin recruitment, defined as the fraction of CTCF sites belonging to each evolutionary/tissue-specific type that falls within a 2-subunit cohesin-bound region.

4.2.6 Cohesin-and-CTCF motif analysis.

Motif identification in CTCF binding sites (both *cis*/*trans*-influenced and evolutionary common/specific) was done using the MEME suite v.5.05[508, 509]. FASTA sequences from the CTCF binding sites were obtained using BEDTools getfasta 2.2.5.0. These sequences were then scanned for CTCF canonical binding motif (M1) JASPAR database (JASPAR motif MA0139.1) using the MEME suite motif scanning function Find Individual Motif Occurrences (FIMO) with default parameters. We used FIMO-assigned CTCF motif orientation and motif scores for further downstream analysis.

In the case of CTCF binding sites with more than one instance of the M1 motif in their sequences, the motif that is closest to the summit of the replicate where the peak signal was at its highest, as defined by the output of peak-calling step using MACS, was selected in the case of evolutionary common/specific sites. The motif with the highest motif score was selected in the case of *cis*/*trans*-influenced CTCF sites because of the variation in SNVs distances to the summit within the binding site. CTCF sites with 0 motif instance in the BL6-specific set of CTCF sites were subject to further motif scanning using alternative CTCF motifs, retrieved from CTCFBSDB 2.0[634, 635] (http://insulatordb.uthsc.edu/download/CTCFBSDB_PWM.mat). Visualisation of both the canonical motif and the alternative motifs was performed using the position weight matrices for each motif as obtained from their respective sources, and carried out using the PWMScan from PWMTools[636].

A motif with a (+) strand orientation indicated that the motif is present on the Watson strand (the + genomic strand), and a motif orientation with (-) orientation is on the Crick strand. We next determined the distance from each cohesin-and-CTCF site with an M1 motif to its nearest cohesin-and-CTCF site. All sites were sorted by chromosome and then by start position, before calculating their distance to the nearest

downstream sites using BEDTools closest 2.2.5.0, with the options -D ref and -t first. The motif combinations between pairs of nearest CTCF sites were defined as "Tandem" if their orientations were in agreement (++/--). If the pair of sites were in opposite orientation, a "Convergent" combination resulted when the upstream motif was in (+) orientation, or "Divergent" combination when the upstream motif was in (-) orientation. The expected proportions of cohesin-and-CTCF pairs of nearest sites based on their evolutionary/tissue-specificity type were calculated from their overall proportion of the total set of sites. For example, the expected frequency of *musculus*-common sites, which constitute 89% of all cohesin-and-CTCF sites, to be nearest to similar *musculus*-common sites if randomly distributed equals 0.89x0.89 = 0.797 (79.7%). The Circos plot from Figure 4.6g was generated using the table visualisation webpage: http://mkweb.bcgsc.ca/tableviewer/[637]

4.3 Results

4.3.1 Depletion of repeat content in *cis/trans*-influenced CTCF sites indicates older evolutionary origin

Having established that the evolution of CTCF occupancy in subspecies-specific manner occurs via the expansion of SINE B2 elements in the short evolutionary time since the divergence of BL6 and CAST (see section 2.3.2), we investigated the contribution of repeat elements activity in driving CTCF occupancy in sites under *cis/trans* binding variation. Analysis of the extent of repeat content enrichment in DNA sequences occupied by TF binding sites influenced by *cis/trans* variation revealed a number of interesting observations of the differences between CTCF and other liver-specific transcription factors.

Compared to the four most represented transposable elements (TE) superfamilies in the mouse genome, CTCF is strongly enriched with SINE TEs; 60%, three times as many as the background genomic sequences masked by TEs ($\chi 2$ test without Yates correction, p-value < 2.2e-16). On the other hand, CTCF is depleted for longer repeat elements, namely LINE and LTR TEs, even though LTRs are second most common TE in *cis/trans* CTCF binding sites, 21% compared to 29% in background ($\chi 2$ test without Yates correction, p-value < 2.2e-16). There is a slight enrichment for DNA-transposons compared to the background, but it only accounted for 5% of the all sequences masked by TEs (Figure 4.1a). The three other TFs (CEBPA, FOXA1 and HNF4A) have higher than expected SINE elements content (1.5-2 times as much as background), with similar depletion in LINE TEs. They do exhibit an elevated enrichment of LTR TEs, as much as SINE TEs contribution to

their sequence (35%-39% for the 3 TFs compared to 29% background). This is consistent with the observed relative enrichment (compared to *musculus*-common) of LTRs in the evolutionarily-young, subspecies-specific binding sites for two of these TFs (CEBPA and FOXA1) in both BL6 and CAST (see section 2.3.2). Similar enrichment for DNA transposons to that of CTCF were also observed for the other TFs, albeit contributing only 6-8% of the all sequences masked by TEs. All differences were statistically significant with $\chi 2$ test and p-values < 2.2e-16.

A breakdown of CTCF sites by their *cis/trans* regulatory categories reveals the same pattern of contribution of the four superfamilies of TEs in all categories, with SINE TEs being the most represented, making up 60% of all repeat content in binding site sequences (Figure 4.1a). No statistically significant differences between either the four categories or the overall pattern for *cis/trans* CTCF sites were found ($\chi 2$ test, p-value > 0.7). *Cis/trans* patterns for the other TFs were also found to reflect the general pattern of the total binding sites for each TF (see Appendix 3, Figure S3.1).

Comparison of the overall repeat content in all binding site sequences between *cis/trans* and non-*cis/trans* sites show that CTCF is the only TF whose *cis/trans* sites are significantly depleted for repeat elements (30% reduction from 29% to 19%), along with a corresponding depletion in SINEs 11% versus 19% in their non-*cis/trans* counterparts ($\chi 2$ test, p-value < 0.05) (Figure 4.1b). Surprisingly, they have slightly higher enrichment with LINE and LTR TEs compared to non-*cis/trans*. Conversely, *cis/trans* in CEBPA and HNF4A have twice the repeat content as their non-*cis/trans* sites ($\chi 2$ test, p-value < 0.01). There was no statistically significant difference between the two with FOXA1 ($\chi 2$ test, p-value > 0.6). The repeat content of the total sequence of the mouse genomic background was found to be higher than both *cis/trans* and non-*cis/trans* TFs binding sites ($\chi 2$ test with Bonferroni correction, p-value < 0.001 for both *cis/trans* and non-*cis/trans* sites across TFs) (Figure 4.1b).

Cis/trans sites depletion in repeat elements is further supported by looking at the proportion of sequence occupied by TEs in the four regulatory categories compared to non-*cis/trans* and randomised genomic regions of the same size. The same overall depletion in repeat elements is seen across the four major TE superfamilies (Figure 4.1c). Significantly lower fraction of the sequences is occupied by SINE elements in *cis/trans* sites than their non-*cis/trans* counterparts (Mann-Whitney U test, p-value < 2.2e-16), and their contribution to sequence is similar to randomised regions level (Mann-Whitney U test, p-value = 0.111). Both *cis/trans* and non-*cis/trans* CTCF sites are depleted for LINEs when compared to random regions (Mann-Whitney U test, p-value < 2.2e-16). This is consistent with recent reports of selection against long sequence deletions at the CTCF sites[638]. All regions studied were enriched for DNA-

transposon over the genomic background, even though their overall contribution to the total sequence occupied by TEs remained marginal (~5%).

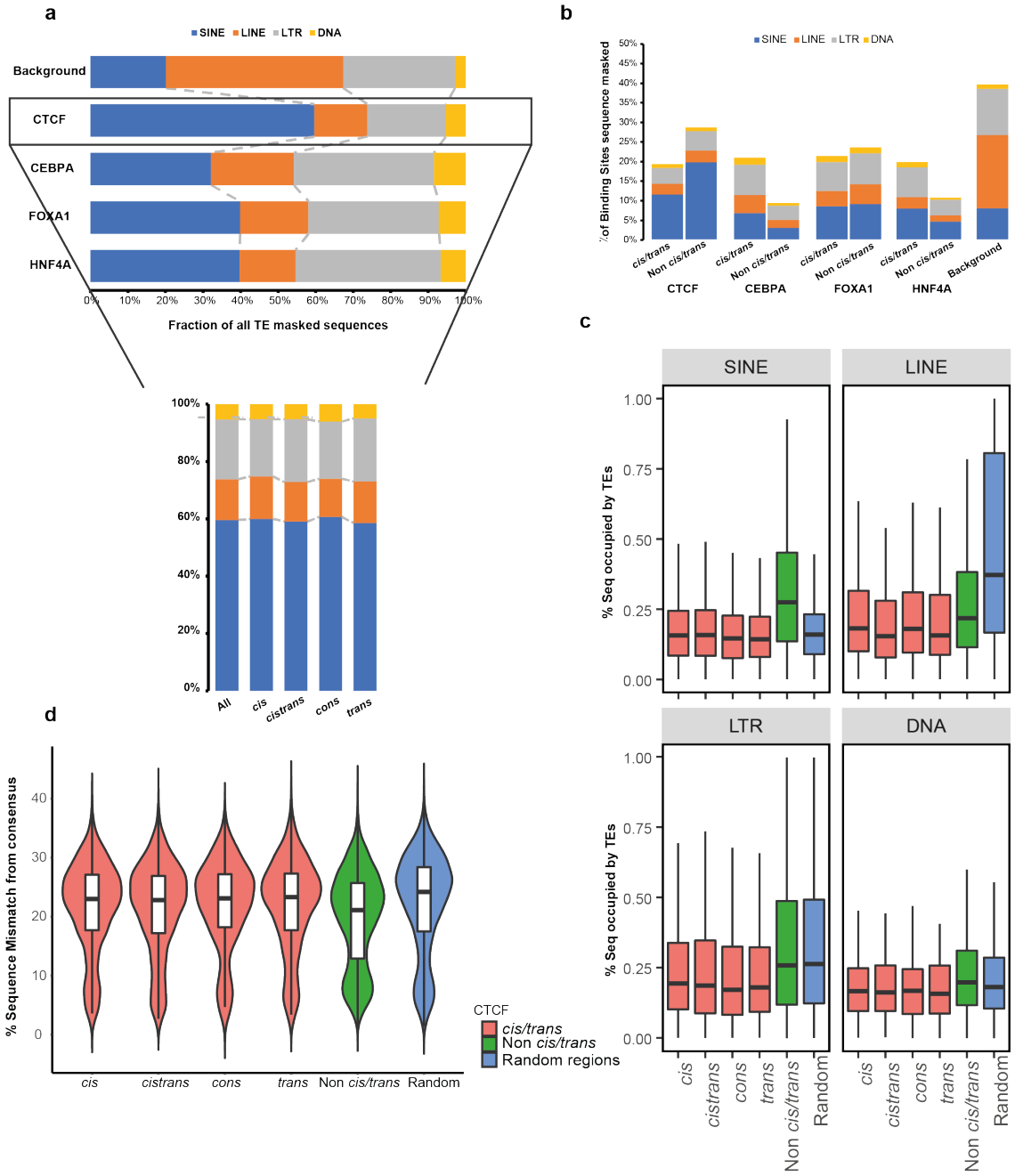


Figure 4.1: *Cis/trans*-influenced CTCF sites are depleted for repeat content

a Horizontal bar chart shows the fractions of different TE superfamilies in *cis/trans* TFs binding sites masked by repetitive sequences. The top bar refers to the percentage each TE superfamily comprises in all repeat masked sequences in the BL6 mouse genome as a background. The vertical bar chart (*below*)

visualises a breakdown of CTCF sites per their *cis*/*trans* category constituents. **b** Stacked bar chart illustrating the extent of sequence masking as a fraction of the total length of sequence occupied by TF binding sites in both *cis*/*trans* and non-*cis*/*trans* assigned peaks. The last bar on the right represents the proportion of repeat content in the total genomic sequence of the BL6 mouse by the type of TE superfamily these repeats originate from. **c** Boxplots displaying the percentage of sequence length of *cis*/*trans* CTCF binding sites that is occupied by TE superfamilies, compared with their non-*cis*/*trans* counterparts and randomised genomic regions matched for length. **d** Violin plot of the fraction of sequence mismatches/substitution from the TE consensus sequence of the SINE TEs in the same categories from **c**. The boxplots within each violin plot show the variation in sequence mismatch level from the consensus in all categories.

The percentage of mismatch in TE sequence from the consensus can be construed as a measure of the relative age of the TE element. As the TE element gets older, the TE consensus sequence acquires more mutations/substitutions in sequence. Evolutionary young sites brought upon via repeat expansion of TE elements have only had a relatively short evolutionary time to accumulate mutations, and hence their mutational load is light. Therefore, binding sites in TE elements characterised by increased level of mismatches in the sequence are considerably older than their mismatch-free counterparts. We used this observation to investigate the relative age of CTCF *cis*/*trans* sites, in comparison with non-*cis*/*trans* sites and matched randomised regions in the most represented TE superfamily, SINEs (Figure 4.1d). The results obtained revealed that CTCF *cis*/*trans* sequences occupied by SINE TEs exhibit a higher degree of mismatches from TE consensus sequences invariably across all four *cis*/*trans* categories. This indicates the significantly longer evolutionary age of these sites in comparison with either non-*cis*/*trans* or randomised regions (Mann-Whitney U test, p-value $<$ 7.517e-15). Since *cis*/*trans* sites represent, by definition, genomic loci where binding CTCF binding is observed between the two subspecies (lineage-specific binding is rare. See Chapter 3), it is likely that these sites are evolutionarily older, and may even further be more deeply conserved through the murine lineage.

4.3.2 Evidence of regulatory potential of *cis*/*trans* CTCF sites at proximal active regulatory elements

In order to establish the potential regulatory effect of *cis*/*trans*-influenced CTCF binding, we evaluated the genome-wide distribution of these sites with relation to the genomic features they occupy. A higher fraction of *cis*/*trans* CTCF sites are found in intergenic regions than any other TF, whereas their binding in intronic regions is distinctly smaller than in liver-specific TFs (Figure 4.2a *left*). A comparatively higher

number of *cis*/*trans* sites are also bound at or around TSSs than the other three TFs. All differences were found to be statistically significantly different between CTCF and other liver-specific TFs ($\chi 2$ test, p-value < 2.2e-16).

CTCF *cis*/*trans* sites overlap exons from 170 genes. Gene Ontology analysis of these genes returned no statistically significant hits for either biological process, molecular function or cellular component, suggesting that this binding is just a reflection of CTCF genomic distribution and is not associated with a particular function. Further investigations of the distribution of the four *cis*/*trans* categories in genomic features revealed the same pattern of distribution to the one observed in all *cis*/*trans* sites, with the vast majority of CTCF sites binding in intergenic/intronic sequences (Figure 4.2a *right*). $\chi 2$ test for differences between *cis*/*trans* categories in their genomic features did not return any statistically significant differences (p-value = 0.25). *Cis*/*trans* patterns for the other TFs were also found to reflect the general pattern of the total binding sites for each TF, although an apparent enrichment in promoter sequences is observed across all liver-specific TFs (see Appendix 3, Figure S3.2).

We next investigated the binding of *cis*/*trans* CTCF and other TFs in/around active regulatory elements. We mapped the genome-wide co-localisation of TF binding sites with genomic location characterised by the presence of the H3K4me3 histone modification, a known promoter marker, and the H3K27ac histone modification, a marker of regulatory enhancer activity, in addition to marking active promoters when coupled with H3K4me3. Results show that, unlike liver-specific TFs, *cis*/*trans* CTCF sites co-binding with markers of gene expression is not common and is generally unfavoured (Figure 4.2b). The effect is most apparent in enhancers, where only 5% of *cis*/trans CTCF sites bind to enhancer sequences, compared to 28-38% in other TFs. Similar pattern is observed with promoter sequences, albeit smaller in scale (10% versus 17-22%) (Figure 4.2b). These differences in regulatory element enrichment across promoters and enhancers were found to be strongly statistically significant ($\chi 2$ test with Bonferroni correction, all p-values < 2.2e-16). These difference reflect the biology of these different factors. CEBPA, FOXA1 and HNF4A are tissue-specific TFs that bind to active regulatory element to activate tissue-specific gene expression[362].

A closer inspection of TF regulatory element occupancy affected by *cis*/*trans* variation across the liver-specific TFs considered in this study reveal that these differences also vary in effect depending on the type of variant present in each binding site. For liver-specific TFs, both *cis*- and *cistrans*-acting variation in the three TFs were underrepresented at promoter regions compared to their proportion of all *cis*/*trans* sites, except for FOXA1 (Binomial test with Bonferroni correction, all p-values < 0.001)(Figure 4.2c). The same was observed in *cis*/*trans* TF binding sites at

enhancer regions, but to a slightly lesser significance. *Cons*-variation acting on TF binding sites was, on the other hand, universally overrepresented for both promoters and enhancers (Binomial-test with Bonferroni correction, p-values < 0.0001) (Figure 4.2c), confirming the results we obtained from looking at the pattern of genomic distribution (see Appendix 3, Figure S3.2). This provides evidence that these *cons*-acting liver-specific TFs do not only bind at regions designated as promoters in annotation, but also marked for regulatory activity with histone modifications in the tissue of interest, exhibiting likely transcriptional potential. *Cis/trans*-influenced CTCF occupancy of promoters/enhancers, on the other hand, does not vary from one type of regulatory category to the other.

When the distance from each of the *cis/trans* CTCF sites to their nearest regulatory element (promoter/enhancer) was measured, we found that the majority of these sites occupy sequences located significantly closer to active regulatory elements. The median distance from *cis/trans* sites to their nearest promoter was 23 kb (16 kb when sites *overlapping* promoter sequences were included). The distances from CTCF *cis/trans* sites to their nearest enhancer was even closer, at a median distance of 17.5 kb (12 kb when sites *overlapping* enhancer sequences were included). *Cis/trans*-acting variants in CTCF binding sites are significantly closer to enhancer elements than promoters (Mann-Whitney U test, p-value $< 2.2e\text{-}16$) (Figure 4.2d). 50% of all non-enhancer overlapping *cis/trans* CTCF sites are within 12 kb of an enhancer, whereas 50% of all non-promoter overlapping sites are within 17 kb of a promoter. 95% of all *cis/trans* CTCF sites are within 75 kb of an active regulatory element (Figure 4.2d). Even though *cis/trans* CTCF binding sites do not appear to be bound at active markers of regulatory activity, they still bind very closely by. This indicates the potential for these CTCF sites to take part in modulating gene expression in cooperation with these active regulatory elements.

CTCF binding sites at proximal active regulatory regions do not exhibit variation in their occupancy pattern based on their *cis/trans* variation status (Figure 4.2e). All four categories of *cis/trans* variation in CTCF occupancy are found at the same distance from active enhancer and promoter sites (Kruskal-Wallis test, p-value $= 0.9914$). Similar to the patterns observed above, there are significantly more CTCF sites in the proximity of enhancer elements than promoters, as evidenced by the increased density of all four type of *cis/trans* variation around enhancer elements (Figure 4.2e). There are, however, no statistically significant difference between any of these categories in terms of their enrichment around either promoters (Kruskal-Wallis test, p-value $= 0.6356$) or enhancers (Kruskal-Wallis test, p-value $= 0.1631$) as their signal signature is identical, strongly peaking in and around the proximity of active regulatory elements.

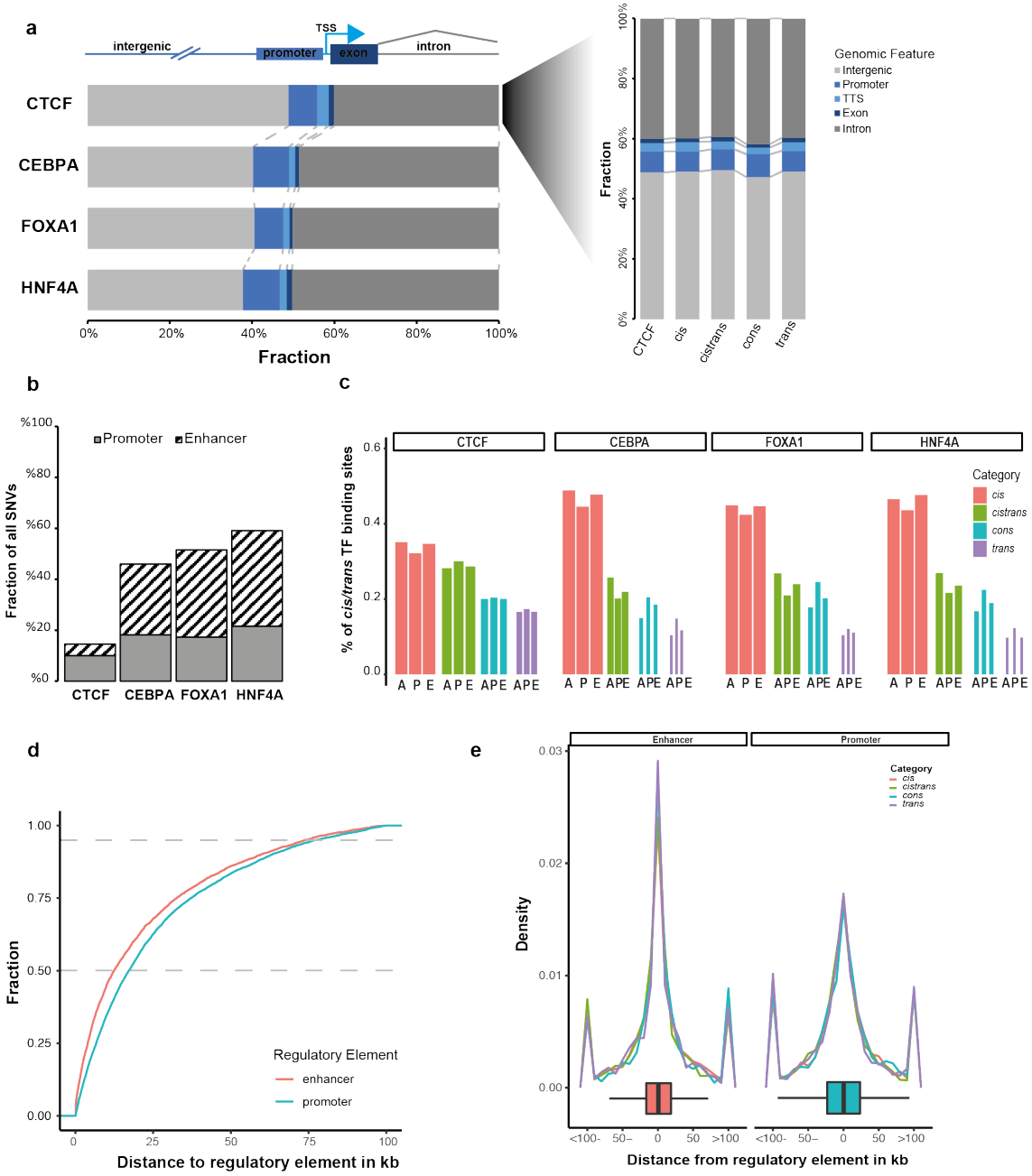


Figure 4.2: Enrichment of *cis*/*trans* CTCF sites at proximal active regulatory elements suggest potential regulatory activity

a Top: a schematic diagram of the different genomic features in which CTCF and other TFs occupancy was measured. The horizontal bar chart highlights the fraction at which different TFs are found in the respective genomic features derived from the most recent mouse genome assembly (GENECODE Release M23 (GRCm38.p6)[629]) . The vertical bar chart shows a breakdown of CTCF

sites genomic features by the type of *cis/trans* variant present at the binding site. **b** Bar chart displaying the fraction of TF sites bound to active regulatory elements, based on co-localisation with histone modifications (see Methods). **c** Proportional bar charts illustrating the fraction of TF locations at promoters (P) and enhancers (E), broken down by their type of regulatory variant present at the binding sites. The widths of bars are based on the overall proportion of binding sites in each category (A). The fractions of *cis/trans* categories localisation in promoters (P) and enhancers (E) for the liver-specific TFs were derived from a random selection of two replicates for each TF. **d** A graph of the empirical cumulative density function for the distance between a *cis/trans* CTCF site and its nearest, non-overlapping regulatory element. The two horizontal dashed grey lines indicate the fraction at which 50% and 95% of all *cis/trans* CTCF sites are at in relation to their distance to the nearest regulatory element defined by the histone modifications present**. e** Density plots of the distribution of CTCF site according to their distance to the nearest non-overlapping regulatory element, by their *cis/trans* category. The boxplots under each curve show the variation in distance toward their respective regulatory element.

4.3.3 Evolutionary young CTCF binding exhibit the same genomic profile of conserved sites

We next looked at whether evolutionary young CTCF sites show hallmarks of possible functionality though binding in the vicinity of active regulatory elements similar to the set of CTCF sites in BL6 and CAST characterised by subspecies-specific SNVs with *cis/trans* binding variation. To establish the potential for functional impact of subspecies-specific CTCF binding sites, we examined the genome-wide distribution of these sites with relation to the genomic characteristics of their occupancy. Consistent with previous estimations[639], 41 to 44% of CTCF binding within both *musculus*-common and BL6-specific sites, respectively, is intergenic, and the remainder takes place within promoters and genebodies (where it occurs mostly in intronic sequences) (Figure 4.3a). Only marginal differences were found between the two categories of evolutionary conservation, with the BL6-specific sites 50% depleted in exonic sequences (with corresponding gains in intergenic and intronic sequences), compared to their *musculus*-common counterparts (Figure 4.3a).

To probe further into the potential for functional roles of these evolutionarily distinct CTCF binding classes, we examined the proximity of these *musculus*-common CTCF sites and subspecies-specific sites, either tissue-restricted or tissue-shared, to transcription start sites (TSSs), calculating the distance from each CTCF binding site to the transcriptions start site (TSS) of the nearest downstream gene. We observed a large proportion of sites near the TSS (median = -11 kb), regardless of the type

evolutionary status or tissue-sharedness of the binding site (Figure 4.3b). The majority of the remaining sites lie further away from the TSS, more than 100 kb upstream. CTCF binding also appears to be depleted directly downstream of the TSS within the gene body. Further analysis looked at CTCF binding site positions relative to the nearest downstream gene. Results revealed an almost identical genomic location distribution, regardless of evolutionary conservation, with the largest portion of CTCF binding more than 10 kb from the nearest downstream genes (Figure 4.3c).

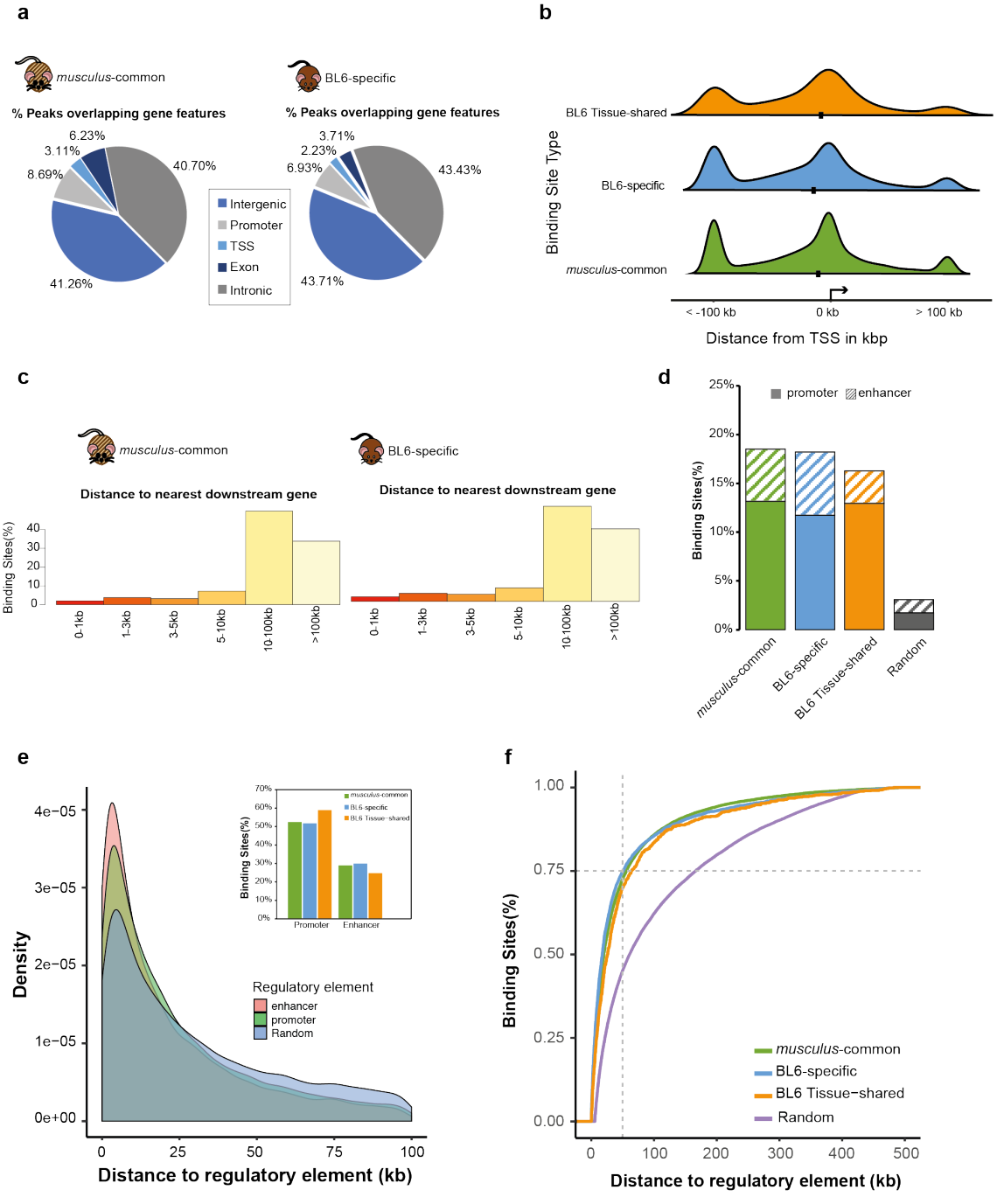


Figure 4.3: BL6-specific binding shares the same characteristics of *musculus*-common CTCF binding.

a Pie chart of the fraction at which *musculus*-common (shared with CAST) and BL6 subspecies-specific CTCF sites are found within the most common gene features (intergenic, promoters, TSSs, exons and intronic). **b** Density plot of the position of CTCF binding sites in terms of their evolutionary/tissue-specificity based on their distance to the transcription start site (TSS) of the nearest downstream gene. The black square inset display the median point of the data. **c** Bar plots of the fraction of CTCF sites in (**a**) in incremental distances to their nearest downstream genes. **d** Bar chart displaying the proportion of CTCF sites in terms of their evolutionary/tissue-specificity at active promoters (H3K4me3 or H3K4me3+H3K27ac) and enhancers (H3K27ac only) against a matched set of random, non-overlapping genomic regions. **e** Density plot of the distance from *musculus*-common and BL6-specific CTCF sites to their nearest, non-overlapping active regulatory region in a ±100 kb window from active regulatory elements, separated by the type of regulatory element compared to the distribution of random genomic regions around the same elements. The bar chart inset shows the type of regulatory element that is closest (but non-overlapping) to *musculus*-common and BL6-specific CTCF sites. **f** Empirical cumulative density function plot for the distance between CTCF binding sites and their nearest, non-overlapping regulatory element, separated based on their evolution/cell-type specificity. The horizontal dashed grey line indicates the fraction at which 75% of all CTCF sites are at in relation to their distance to the nearest regulatory element, with the vertical marking that distance to 50kb. The purple line indicates the distance from the random set of regions (**e**) to their nearest non-overlapping regulatory element.

The proximity of CTCF sites to regions upstream of TSS indicate that these sites may elicit their functional impact by acting on/collaborating with active histone modifications near these regions of regulatory activity. We mapped the genome-wide co-localisation of all CTCF binding sites with genomic locations characterised by the presence of the H3K4me3 and H3K27ac histone modifications. As observed previously with *cis*/*trans*-influenced CTCF sites, results revealed that, regardless of evolutionary class or tissue-specificity, CTCF occupancy does not often coincide with markers of gene expression (Figure 4.3d). Only 12-13% of all CTCF binding sites bind within promoter sequences (H3K4me3), and a further 3-6% co-bind in regions characterised by the presence of the H3K27ac histone modification alone, marking active enhancers. BL6 tissue-shared sites appear to be slightly depleted in enhancers, but these differences were not found to be significant ($\chi 2$ test, p-value > 0.05). Despite marginal CTCF occupancy within these regulatory elements, it is still significantly higher than

would be expected for a chromosome-matched, non-overlapping set of genomic regions of the same length and number (Binomial test with Bonferroni correction, all p-values < 0.001).

Although CTCF sites do not show substantial co-occupancy within active enhancers/promoters, they could still bind sufficiently close to these elements to allow them to play a role in gene regulation. We determined the distance from each of these CTCF sites to their nearest, non-overlapping active regulatory element using the two markers above. We found that the majority of these sites occupy sequences located significantly closer to active regulatory elements, with median distances of 22.7 kb and 17.7 kb to their nearest promoter and enhancer, compared to 105.6 kb and 75 kb, respectively, for a matched set of randomised genomic sequences (Figure 4.3e). This statistically significant difference in proximity between promoters and enhancers (Man Whitney U test, p-value = 0.02) could be explained by the enhancers greater genome-wide number, making it more likely for a CTCF site to be nearer to an enhancer than a promoter. On the other hand, promoters are the nearest regulatory element to 52% of CTCF binding sites, both *musculus*-common and BL6-specific (59% in BL6 tissue-shared sites). Enhancers are the closest to CTCF in 29-30% of sites, with slightly fewer (25%) BL6 tissue-shared sites near enhancers (4.3E *bar chart*).

Stratifying the distance to the nearest regulatory element by the evolutionary class and tissue-specificity of CTCF binding revealed that BL6 tissue-shared sites lie the furthest away from active regulatory regions (median distance = 30.2 kb) compared with either *musculus*-common or all BL6-specific sites (20.9 and 19.8 kb respectively). Differences between these categories; however, were not large enough to be deemed significant in either promoter (Kruskal-Wallis test, p-value = 0.6552) or enhancer (Kruskal-Wallis test, p-value = 0.9451). 75% of all CTCF sites were found within a 50 kb distance of regulatory element (Figure 4.3f). Taken together with the results from 4.3.2, the results indicate the potential for even evolutionarily young, tissue-shared or otherwise, CTCF sites to be involved in regulating gene expression in cooperation with exiting active *cis*-regulatory elements. They also provide evidence that subspecies-specific CTCF binding perform similar functions to the more conserved *musculus*-common sites.

4.3.4 *Cis/trans* CTCF occupancy is strongly TAD-boundary associated

CTCF binding has repeatedly been reported to contribute to the formation and maintenance of topologically-associated domains (TADs), by colocalising with the cohesin protein complex[623, 640]. We next set out to investigate whether regulatory variation in CTCF occupancy brought by *cis*/trans-acting variants in the binding sites

influences the binding of CTCF near TAD boundaries. CTCF binding sites under influence of regulatory variation acting in *cis*/*trans* tend to cluster around TAD boundaries, reaching their highest density just up/down stream from the boundary (Figure 4.4a). The proportion of TAD boundary-associated CTCF sites, defined as CTCF sites bound within a 50kb window of the nearest TAD boundary[575], is higher than that of any of the three other liver-specific TFs (40% compared to 20% only in CEBPA, FOXA1 and HNF4A, binomial test with Bonferroni correction, p-value < 2.2e-16), or their non-*cis*/*trans* counterparts (26%) (binomial test, p-value < 2.2e-16)(Figure 4.4a). The median distance from any *cis*/*trans*-influenced CTCF site to its nearest TAD boundary is 79.6kb, over 30% shorter than the median distance of any liver-specific *cis*/*trans* TF sites (112-116kb). These differences were observed to be statistically significant (Kruskal-Wallis test, p-value < 2.2e-16).

A breakdown of *cis*/*trans* CTCF sites distribution with regards to their proximity to TAD boundaries revealed that there is no observed difference between any of the four regulatory categories of CTCF binding and their clustering around the TADs. About 39-40% of all *cis*/*trans* CTCF sites were found to be TAD-boundary associated, regardless of their regulatory category, and their overall frequencies reflected their general contribution to the total number of CTCF sites, with *cis* and *cistrans* sites being the most abundant (Figure 4.4b *top*). There was no statistically significant difference in the distribution of these sites near TAD boundaries ($\chi 2$ test without Yates correction, p-value > 0.05). A closer inspection of the immediate neighbourhood (10kb up/downstream) of the TAD boundary for the regulatory composition of CTCF binding sites confirmed the previous observation. All four categories are proportionally bound in numbers reflecting their contribution to the total pool of *cis*/trans-acting variants on CTCF binding (Figure 4.4b *bottom*). No statistically significant differences were observed between categories bound in 1kb incremental bins around the TAD boundary ($\chi 2$ test without Yates correction, p-value = 0.78)

We next evaluated the abundance of TF binding sites under *cis*/*trans* acting variation in TADs to understand the pattern of their genomic distribution and the possibility of clustering in regulatory modules. Out of a total of 3643 mouse liver derived TADs, 368 (10%) were free from any CTCF binding, fewer than any of the three other liver-specific TFs (17%-25%) (Figure 4.4c). However, this parity is significantly reduced in TADs harbouring a single TF binding site. TADs containing more than 1 TF binding sites are strongly enriched for the binding of CTCF in comparison with other TFs, although this effect decreases as more TF sites are bound in TADs until this difference is diminished (at about 7-9 TF binding sites per TAD). This trend is reversed for all TADs with 10 or more TF binding sites, wherein liver-specific TF occupancy becomes the most abundant (Figure 4.4c). All these differences,

except in the case of TADs with 7-to-9 sites, were found to be statistically significant ($\chi 2$ test with Bonferroni correction, p-value < 0.0001).

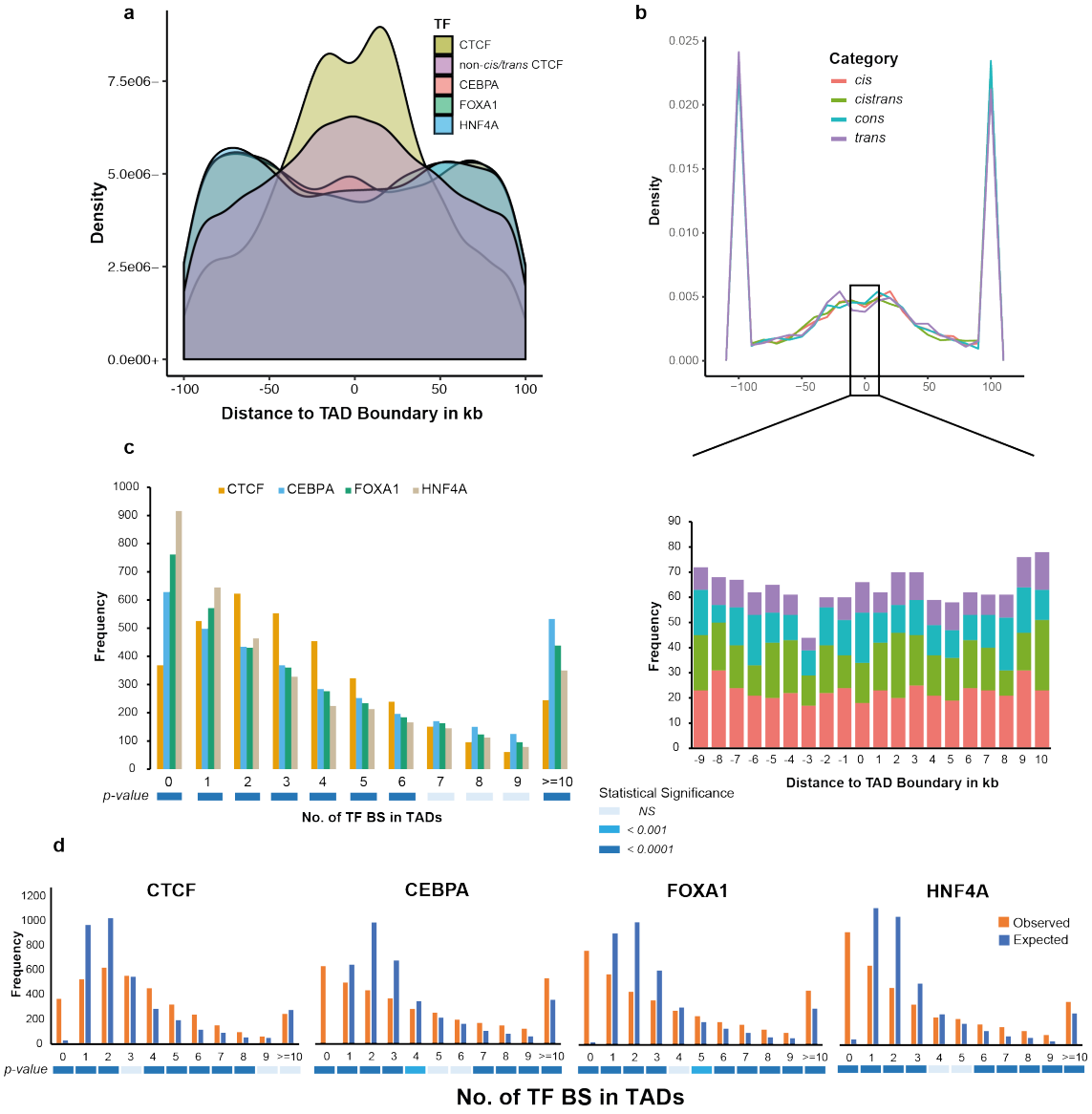


Figure 4.4: *Cis/trans* CTCF occupancy is strongly TAD-boundary associated

a Density plots illustrating the relative enrichment of *cis/trans* TF binding sites within a ±100 kb window up/downstream from the TAD boundaries. The non-cis/trans CTCF sites also shown to illustrate the difference in TAD-boundary association between CTCF sites with SNVs and those without. **b** Frequency polygons (Top) showing the density of four different categories of *cis/trans* variation in CTCF binding sites around TAD boundary. The higher spikes in density at ±100 kb from the TAD boundaries are the results of pooling together

all sites at a distance of more than 100 kb up/downstream from the nearest TAD boundary. The stacked bar chart (Bottom) represents the absolute frequencies of the different *cis*/*trans* categories of CTCF sites at 1 kb intervals of the TAD boundary. **c** Bar chart of the number of *cis*/*trans* TF binding sites found in individual TADs. The counts for all TADs with 10 or more TF binding sites were pooled together. The strip (*below*) denotes the level of statistical significance for the comparison between all TFs (all χ^2 test with Bonferroni correction, p-values > 0.004 considered not significant (NS)). **d** Bar chart of the number of observed *cis*/trans TF binding sites (**c**) compared to the number of sites expected under the assumption of random distribution in single TADs. The counts for all TADs with 10 or more TF binding sites were pooled together. The strip (*below*) denotes the level of statistical significance for the comparison between observed and expected counts (all binomial test, with Bonferroni correction, p-values $>$ 0.001 considered not significant (NS)).

These differences in TF binding in TADs were additionally found to be either more (or less) than expected based on the assumption of random distribution in genomic sequences and TAD sizes (Figure 4.4d). There were universally more TADs with no TF binding for any of the four factors than would be expected, and significantly fewer than expected in 1-3 per TAD. As mentioned earlier, CTCF was more abundant in smaller numbers per TAD (4-6) than either expected or compared to the other TFs. In TADs with 10 or more TF binding sites, liver-specific TFs are found in higher abundance than expected, whilst CTCF frequencies are within the range of their background genomic distribution.

This indicates that CTCF, a tissue-wide factor with highly conserved binding across cell-types, tend to bind across the genomic landscape of TADs, which in turn dilutes the number of sites found per TAD. Tissue-specific TF occupancy, on the other hand, favour clustering within TADs that harbour the regulatory modules they bind to activate gene expression.

4.3.5 Evolutionary young, tissue-shared binding actively associates with cohesin.

The strong association observed between *cis*/*trans* CTCF sites with TAD boundaries suggests these sites contribute to the regulation of chromatin folding and overall genomic architecture. To investigate whether evolutionary young sites exhibit the same preference to TAD-boundary association as their subspecies-orthologous counterparts, we next evaluated the possible contribution of different evolutionary classes of CTCF to large-scale 3D genome structure. We leveraged available HiC experiments that determined the position of topologically-associated domain (TAD)

boundaries in liver[238] to determine the distribution of each type of CTCF binding site to the nearest TAD boundary. Apart from known CTCF enrichment around the TAD boundaries, the majority of CTCF binding sites are located well inside the TADs and further away (> 100 kb) from the boundaries themselves (Figure 4.5a).

At all distances from the TAD boundary, both the frequency and the approximate ratio of *musculus*-common, BL-specific and tissue-shared CTCF binding sites are the same (Figure 4.5a *bar plot*). This ratio holds also true in 1 kb increments from the TAD boundaries. The *musculus*-common sites are expectedly[312] enriched around the TAD boundaries compared to their subspecies-specific counterparts whether or not they are tissue-shared (Figure 4.5a). Thus, tissue-dependence of BL6-specific sites does not affect their distribution around the TAD boundaries. These results suggest that although TADs do not vary across tissues[235], some TAD boundaries may, at least partially, be maintained by tissue-specific CTCF binding.

Increased occupancy of CTCF sites around TAD boundaries is reportedly accompanied by co-localising with cohesion-protein complex to form chromatin loops[80, 641]. CTCF-associated cohesin ring formation is known to be instrumental in chromatin loop formation and maintenance[219]. Upon binding to the two ends of a transcription regulatory unit, CTCF recruits cohesin to aid in the formation of a chromatin loop (Figure 4.5b *diagram*). To elucidate potential functional involvement of these CTCF elements, we determined the level of co-location of CTCF with cohesin complex proteins in BL6 mice. This was done using ChIP-seq data from adult mouse liver in two biological replicates for three proteins from the cohesin complex: RAD21, STAG1 and STAG2[362].

Co-location of CTCF and cohesin was determined using a minimum stringent condition of at least two cohesin subunits whose binding overlaps the same genomic segment binding CTCF. All classes of CTCF binding sites show cohesin co-location with the highest level (approximately 80%) observed for *musculus*-common CTCF sites. BL6-specific CTCF sites co-localised with a significantly reduced level of cohesin such that only half of these sites showed signal from at least two cohesin subunits (p-value = 2.8×10^{-9}). However, of those BL6-specific sites, the tissue-shared subset (i.e. the ones bound in all five tissues in Figure 2.4g) exhibited cohesin co-localisation at essentially the same level as the set of all CTCF binding sites, and only slightly less than that of the *musculus*-common sites (Figure 4.5b).

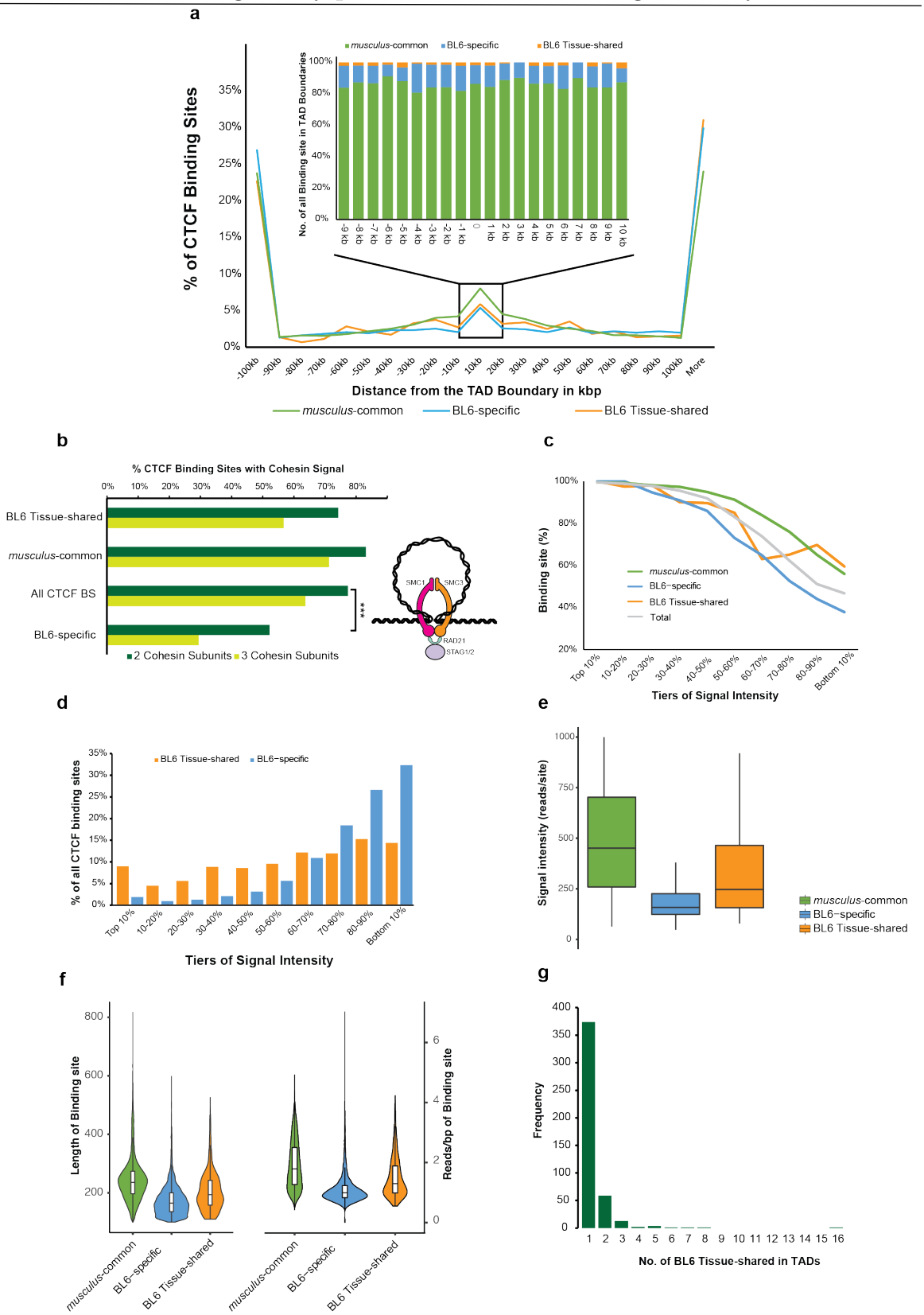

a
musculus-common
BL6-specific
BL6 Tissue-shared
% of CTCF Binding Sites
No. of all Binding site in TAD Boundaries
Distance from the TAD Boundary in kbp
b
% CTCF Binding Sites with Cohesin Signal
BL6 Tissue-shared
musculus-common
All CTCF BS
BL6-specific
2 Cohesin Subunits
3 Cohesin Subunits
SMC1
SMC3
RAD21
STAG1/2
c
Binding site (%)
Total
Tiers of Signal Intensity
d
% of all CTCF binding sites
Tiers of Signal Intensity
e
Signal intensity (reads/site)
f
Length of Binding site
Reads/bp of Binding site
g
Frequency
No. of BL6 Tissue-shared in TADs

Figure 4.5: Recent BL6 tissue-shared CTCF binding efficiently recruits cohesin and is associated with higher ChIP-signal

a Plot of the distance from each of the CTCF binding sites to the nearest up/downstream topologically-associated domain (TAD) boundary. The outline box indicates the region between -10 kb and +10 kb from the nearest TAD boundary. The bar chart below shows the proportion of CTCF sites in this region in 1 kb intervals based on their evolutionary/tissue-specificity type. **b** Bar chart of the recruitment of cohesin-complex subunits by CTCF. The x-axis shows the percentage of CTCF binding sites in which a cohesin-complex signal was found. The schematic diagram next to the bars is an overview of the structure of the Cohesin-complex. The asterisks indicate the significance of Chi-square goodness-of-fit test (p-value = 2.8x10-9). **c** Line plot of CTCF-cohesin co-occupancy at matched tiers of signal intensity (reads/site) for CTCF binding sites. All categories of CTCF where within a distinct range of signal for each bin, even though their numbers were variable. Co-occupancy was calculated as the number of CTCF sites that co-localise with a 2-subunit cohesin-bound region. **d** Bar plot of the percentage of BL6-specific sites (subspecies-specific and the tissue-shared subset) present within each signal tier from (**c**), from the total set of CTCF sites of that type. **e** Box plot of the variation in signal intensity across the different types of CTCF binding sites. **f** Violin plots of the kernel density of CTCF binding sites according to the length of the binding site (*left*) and the depth of sequencing (read/bp) of the different evolutionary categories of CTCF. **g** Plot of the number of BL6-specific tissue-shared CTCF binding sites in single TADs.

This increased level of cohesin recruitment coinciding with BL6 tissue-shared binding sites in comparison to their tissue-variable counterparts hints at increased involvement in the formation of chromatin loops, and perhaps important functions. The same pattern was also observed for all types of CTCF binding sites with three cohesin subunits instead of two, albeit with reduced ratios (Figure 4.5b *yellow bars*). This could be attributed to the uniformly lower ChIP enrichment from the cohesin subunit STAG1 which limited our sensitivity to detect three cohesin subunits.

We next asked whether the differences in cohesin recruitment in BL6 tissue-specific sites can be explained by lower overall CTCF ChIP signal enrichment of these sites. We determined CTCF cohesin recruitment at matched tier of ChIP signal intensities for all evolutionary/tissue classes of CTCF binding sites. At the top 10% of the signal, almost all CTCF sites are associated with cohesin recruitment to the DNA, regardless of the evolutionary or tissue-specificity of the binding site (Figure 4.5c). However, as signal decreases, CTCF-cohesin co-occupancy markedly decreases, particularly in BL6 tissue-specific sites which start to show accelerated reduced

recruitment at 30-40% of the signal. At the bottom 10% tier of the signal, only 56% of *musculus*-common CTCF sites are associated with cohesin, and even fewer at BL6 tissue-specific sites (38%). Curiously, cohesin recruitment for BL6 tissue-shared sites is generally intermediary between the former two, yet rises considerably at the last two tiers above both the total and *musculus*-common level of recruitment (Figure 4.5c). This indicates that although reduced cohesin with lower ChIP signal is indeed stronger in BL6 tissue-specific, the effect does not quite register for the tissue-shared subset of subspecies-specific sites, remaining generally high even at low signal values. This observation, however, could be the result of the small number of CTCF sites in this particular category, especially at lower tiers of signal intensity.

We further determined the proportions of these BL6-specific sites within each tier of signal intensity. The results revealed that only 2% of these sites (136) were found within the top 10% range of signal (Figure 4.5d), 82 of which (60%) transpired to be tissue-shared. The proportion of BL6-specific sites in subsequent tiers rose gradually, but over half of all subspecies-specific sites occupied the lower 20% range of the signal. On the hand, proportion of BL6-specific tissue-shared sites in the different tier of signal remained roughly the same, and even though they rose slightly towards the lower end of signal, that coincided with higher cohesin recruitment than all other classes of CTCF binding (Figure 4.5c). Furthermore, whilst BL6-specific tissue-shared sites make up over half the subspecies-specific CTCF sites in the top 40% range of the signal, their proportions quickly dwindle at the lower 50% range of the signal, and as the tissue-specific sites become more predominant, levels of cohesin recruitment lower.

This is further evidenced by comparing signal values of CTCF binding across all three classes (Figure 4.5e). BL6 subspecies-specific sites have a significantly lower signal (median = 158 bp/read) than either *musculus*-common (451) and BL6 tissues-shared sites (246.5) (Man Whitney U test, p-values $<$ 2.2e-16 in comparison with both). These results were not affected by the size of the binding site, as both *musculus*-common and tissue-shared sites have longer peak sequences (Man Whitney U test, p-values $<$ 2.2e-16) in spite of which they still exhibited higher depth of sequencing than their tissue-specific equivalents (Figure 4.5f). In sum, these results suggest that BL6 tissue-shared are mostly responsible for heightened cohesin recruitment for subspecies-specific CTCF sites. As the number of their tissue-specific counterparts start to increase at lower levels of ChIP signal, their association with cohesin-complex subunits is reduced.

We investigated whether these BL6 tissue-shared CTCF-cohesin regions may participate in novel looping interactions and found no clear evidence. We speculated that BL6 tissue-shared CTCF-cohesin regions could act as novel loop anchors within established TADs and so limited our analysis to TADs with a minimum of two of these

regions. There were fewer than 100 such regions in the whole genome and more than half of the potential novel loop anchors are found in single sites per TAD (Figure 4.5g). We suggest that these sites are either stabilising existing TAD structures or maybe involved in intra-TAD domain loop formation.

4.3.6 CTCF sites under *cis/trans*-acting variation highly co-localise with cohesin-complex proteins

We followed up the analysis in 4.3.4 with an investigation of TF binding sites co-occupancy with 3 cohesin subunits, Rad21, STAG1 and STAG2, in sites characterised by variation in *cis/trans* in CTCF and the three liver-specific transcriptions factors (see section 4.3.5). The results confirm that CTCF sites under the influence of *cis/trans* regulatory variation strongly co-localised with cohesion-protein complex subunits, with 90% of all sites (>13000) binding in a region where a minimum of two cohesin subunits bind (Figure 4.6a). Slightly fewer sites bind in regions characterised by the binding of 3 cohesin-subunits (slightly over 80%). The enrichment of cohesin-bound regions is distinctly weaker in liver-specific TFs, even though HNF4A does appear to show co-binding with a minimum of 2-subunit cohesin-regions in 50% of its peaks (Figure 4.6a *left*). Liver-specific TF binding in cohesin-complex regions was found to be associated mostly with cohesin-non-CTCF (CNC) sites invariably for all three TFs (Figure 4.6a *right*). This is consistent with previous reports of CNC enrichment at tissue-specific promoters and enhancers that function to maintain TF complexes at regulatory elements [224, 362].

Cis/trans CTCF site bound in cohesin-complex regions exhibit the same pattern of enrichment in the vicinity of TAD boundaries as seen in 4.3.4 (Figure 4.6b). The median distance from these sites to the nearest TAD boundary was found to be 77.8 kb, only marginally shorter than the distance reported above. That is not unexpected as almost all *cis/trans* CTCF sites co-bind with cohesin subunits, thus the pattern of their binding should be comparable. This difference in distance to TAD boundary was found to be statistically insignificant (t-test, p-value = 0.1573). This pattern of enrichment near TAD boundaries invariably matched across the different categories of *cis/trans* CTCF sites, with no category showing any statistically significant difference (Kruskal-Wallis test, p-value = 0.2514) (Figure 4.6b).

The results further indicate that the majority of all *cis/trans* CTCF sites that co-localise with cohesin-complex subunits are TAD-boundary associated (Figure 4.6c). Nearly 75% of sites in ±100 kb window from the TAD boundary are within a distance of 50kb up/downstream from the nearest TAD boundary, regardless of the type of *cis/trans* variation acting on the CTCF binding site. These results strongly suggest a role of these sites in the regulation and maintenance of TAD boundaries.

These CTCF-cohesin occupied regions are found near the base of the loop, whereby a convergent orientation of their motifs is generally preferred during looping. Looping of chromatin whereby the CTCF motifs at the base are in tandem orientation is also observed, albeit in reduced capacity [303]. We first investigated the enrichment of the canonical motif (M1) in *cis/trans*-influenced CTCF binding sites. Nearly 97% of these sites harbour one or more instances of the canonical CTCF motif within the peak sequence (Figure 4.6d). Two thirds of sites have only a single motif within the sequence. For peaks with more than one motif instance, the motif with the highest score/p-value was selected.

In order to establish the orientation of CTCF motifs are the base of possible chromatin loops, we investigated the motif combinations between pairs of CTCF sites that co-localise with cohesion-subunits. We set a minimum distance of 5kb up/downstream from the nearest potential CTCF binding site. We were able to identify 7456 pairs of CTCF sites that met these criteria. The pie chart in Figure 4.6e shows a breakdown of the three possible combinations in these pairs of *cis/trans* CTCF-cohesin sites. Almost half (48%) the motif combinations of pair of *cis/trans* CTCF sites were in tandem orientation (++/--). A third of all pairs had a divergent motif combination pattern of (-+), i.e. the upstream CTCF motif was on the reverse strand, whereas its downstream counterpart was on the forward strand. Convergent motif combinations in neighbouring CTCF sites were found to be slightly in the minority (22%). The differences in motif combinations between the three different types were found to be statistically significant ($\chi 2$ test, p-value 0.011).

Next, we looked at the distance between these CTCF binding sites to their nearest neighbour and the motif combinations they form. Tandem-oriented pair of CTCF motifs and those in divergent orientation were found to have shorter distances to their nearest potential pair than their convergent counterparts, which were found to have the furthest distance from their pairs (Figure 4.6f). These differences were found to be statistically significant between convergent pairs and both tandem and divergent pairs (Man Whitney U tests, p-values < 0.00001), but not between the latter two (p-value $= 0.24$) (Figure 4.6f).

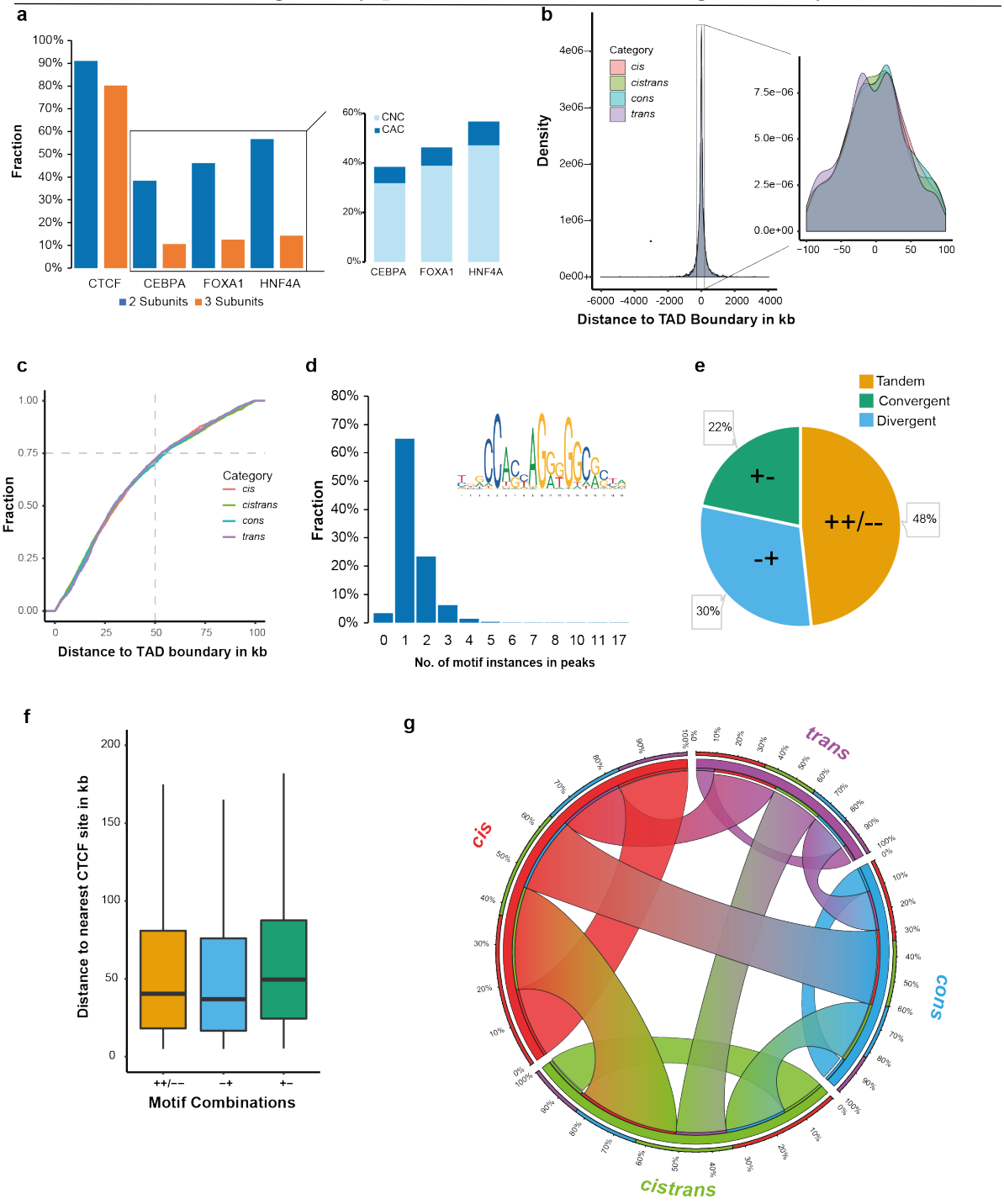


Figure 4.6: TAD-boundary associated *cis*/*trans* CTCF occupancy is accompanied with cohesin-complex co-localisation

a Bar chart (*left*) representing the proportion of *cis*/*trans* TF binding sites co-localising with cohesin-complex regions characterised by the presence of 2-3 subunits. The second bar plot (*right*) shows a breakdown of the type of cohesin-

complex regions liver-specific TF sites (*boxed*) were found to co-localise with (CAC= cohesin-and-CTCF, CNC= cohesin-non-CTCF). **b** Density plot (*left*) showing the distribution of *cis*/trans CTCF-cohesin associated sites in terms of their distance to the nearest TAD boundary. The smaller density plot (*right*) offers a closer look at the ±100 kb window around the TAD boundary of the same distribution, separated by the type of *cis/trans* variant present in the CTCF binding site. **c** Empirical cumulative density function plot for the distance between a cohesin-associated *cis/trans* CTCF site and its nearest TAD boundary. The horizontal dashed grey line indicates the fraction at which 75% of CTCF sites in ±100 kb window from the TAD-boundary are within a ±50 kb distance, the vertical dashed grey line, for the 4 *cis/trans* categories of variation in CTCF sites co-localised with cohesin-complex subunits. **d** Bar chart of the number of instances the CTCF M1 canonical motif (in the panel above the graph area) was identified in the peak sequence of *cis/trans* CTCF sites. **e** Pie chart of the relative proportions of motif combinations between pairs of neighbouring cohesin-associated, *cis/trans*-influenced CTCF sites that are within a minimum distance of 5 kb from each other. **f** Boxplots of the distances of the CTCF sites from **e** their nearest CTCF site in kb by the motif combination they form. **g** Plot of the *cis/trans* categories of each pair of neighbouring (min. distance = 5kb) CTCF binding sites. Ribbon size encodes the fraction of each category's association with the other categories in pairs. The outermost circle shows the relative proportions of each combination for the total number of sites in each category separately.

These differences, however, did not translate to differences in the proportion in the various categories of CTCF binding sites under *cis/trans* variation in terms of their combinations (Figure 4.6g) or their distances to the nearest CTCF sites. The proportions of combinations of CTCF sites reflected their overall proportions, with *cis*- and *cistrans*-variation in CTCF binding being the most common either in combinations within the same category, between them, or with the other two categories forming the majority of pairs of neighbouring sites. Contributions from the other two categories were also equivalent to their overall proportions.

4.3.7 Tissue-shared binding clusters closer to regions of CTCF binding and favours tandem motif orientation.

As explained above, a pair of CTCF bound to their sites in the favourable convergent orientation of their motifs recruit cohesin subunits to the base of the loop, although chromatin loop of tandemly oriented CTCF motifs is also reportedly observed. To determine the potential for the evolutionarily young sites to take part in

such formations, we analysed motif content and orientation of these cohesin-associated CTCF sites.

We scanned *musculus*-common and BL6-specific (tissue-shared and tissue-specific) CTCF sites for presence of the M1 canonical binding motif (Figure 2.1b). Whereas only 5% of all *musculus*-common sites have 0 instances of the motif within their sequence, (9%) did not carry the motif in BL6 tissue-shared sites, and almost 15% of BL6 tissue-specific sites lack the canonical motif in their peak sequence (Figure 4.7a *left*). These differences were statistically significant compared to the *musculus*-common set (Binomial test, p-values 7.131e-07 and $< 2.2e-16$ for tissue-shared and tissue-specific respectively). We next investigated these 0-motif sites for the presence of 5 alternative CTCF motifs (Figure 4.7a *right*) obtained from the CTCFBSDB 2.0 database[642]. 40% of BL6-spcific tissue-shared sites harboured one or more of alternative motifs in their sequences, yet tissue-specific sites again showed a depletion in motif content with only a third of sites having any alternative motifs at all within their CTCF binding sequences (Figure 4.7a *middle*). For all subsequent analysis, the closest canonical motifs to the peak's summit were only selected.

Whereas motif orientation is expectedly invariable among the different categories of CTCF binding (*musculus*-common, BL6-specific tissue-shared and tissue-specific), the distance from the closest motif to the summit significantly differs in BL6 tissue-specific sites, whereby ~54% of sites have their motif further away from the summit (χ^2 test, p-values 1.94926E-78 and 1.14951E-11 compared to *musculus*-common and BL6 tissue-shared respectively) (Figure 4.7b). The increased distance between the summit and the motif from *musculus*-common to BL6 tissue-specific sites coincided with a corresponding drop in motif score (Figure 4.7c). These differences were all found to be statistically significant in both overall and one-to-one comparisons (Kruskal Wallis p-values all $< 2.2e-16$ for motif score and distance. Man Whitney U tests, p-values range from 5.257e-08 to $< 2.2e-16$).

We investigated the motif combinations between pairs of cohesin-associated *musculus*-common and BL6-specific (tissue-shared and tissue-specific) CTCF sites to determine the orientation of CTCF motifs are the base of possible chromatin loops, using a minimum distance of 5kb from the nearest CTCF binding site. Based on these criteria, *musculus*-common sites were found to form pairs with their nearest CTCF sites in the same frequency as expected based on their overall proportion of all CTCF binding sites (Figure 4.7d).

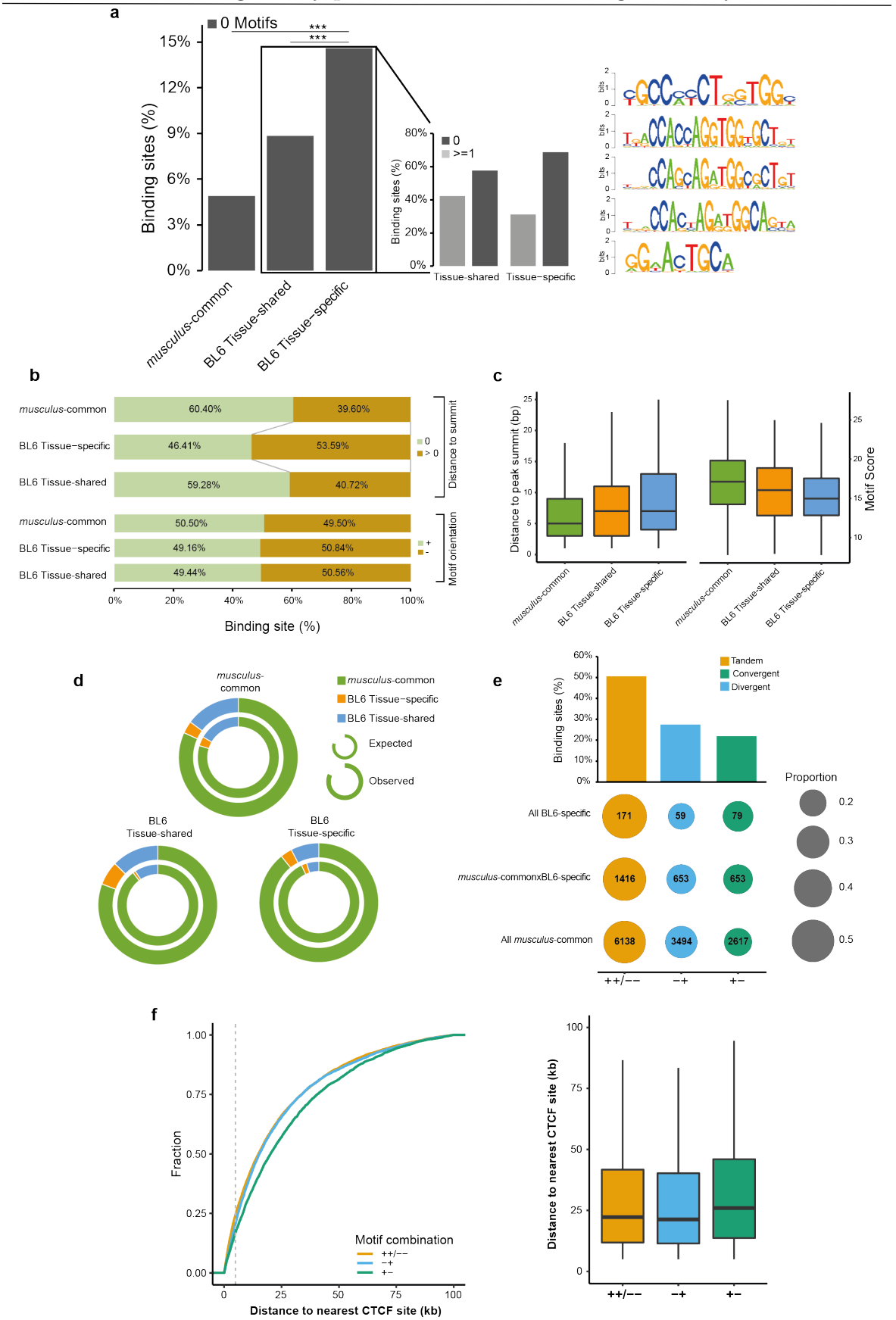

a
0 Motifs
***
***
15%
12%
9%
6%
3%
0%
Binding sites (%)
musculus-common
BL6 Tissue-shared
BL6 Tissue-specific
80%
60%
40%
20%
0%
0
>=1
Tissue-shared
Tissue-specific
b
musculus-common
BL6 Tissue−specific
BL6 Tissue-shared
60.40%
39.60%
46.41%
53.59%
59.28%
40.72%
0
> 0
Distance to summit
50.50%
49.50%
49.16%
50.84%
49.44%
50.56%
+
-
Motif orientation
0%
20%
40%
60%
80%
100%
Binding site (%)
c
Distance to peak summit (bp)
Motif Score
musculus-common
BL6 Tissue-shared
BL6 Tissue-specific
d
musculus-common
musculus-common
BL6 Tissue−specific
BL6 Tissue-shared
Expected
Observed
BL6
Tissue-shared
BL6
Tissue-specific
e
Tandem
Convergent
Divergent
Binding sites (%)
Proportion
All BL6-specific
musculus-commonxBL6-specific
All musculus-common
171
59
79
1416
653
653
6138
3494
2617
++/−−
−+
+−
0.2
0.3
0.4
0.5
f
Fraction
1.00
0.75
0.50
0.25
0.00
Motif combination
++/−−
−+
+−
0
25
50
75
100
Distance to nearest CTCF site (kb)
Distance to nearest CTCF site (kb)
++/−−
−+
+−

Figure 4.7: Motif characteristics of CTCF binding in sites with evolutionary/tissue-specificity variation

a Bar plot (*left*) of the proportions of *musculus*-common and BL6-specific (tissue-shared and tissue-specific) CTCF binding sites lacking the canonical M1 motif in their peak sequence. The asterisks indicate the significance of Chi-square test (p-value < 0.0001) for BL6 Tissue-shared sites versus the other two categories. The smaller bar plot (*middle*) shows the proportion of those 0-motif BL6-specific CTCF sites (*boxed*) in which one of the other CTCF motifs from CTCFBSDB 2.0 (*rightmost*) were alternatively identified. **b** 100% bar plots of the fraction of *musculus*-common and BL6-specific (tissue-shared and tissue-specific) CTCF sites according to the distance of their motifs to the peak summit (*top*) and the motif orientation (*bottom*). **c** Box plots of the distribution of non-overlapping motifs (>0) distance to the peak summit in bp (*left*) and the motif scores as calculated from the information content (PWM) of the canonical motif[508](*right*) according to their evolutionary/cell-type status. **d** Donut charts of the evolutionary/tissue-specificity type of CTCF sites that cohesin-associated CTCF sites are nearest to (min. distance 5 kb). The outer circle indicates the observed proportions for the identities of the other cohesin-associated CTCF sites in the pairs, compared to the expected proportions (*inner circle*). **e** Bar plot (*top*) of the relative proportions of motif combinations between pairs of neighbouring cohesin-and-CTCF sites that are within a minimum distance of 5 kb from each other. The circles (*bottom*) display a breakdown of each motif combination observed based on the evolutionary status of the neighbouring CTCF sites, with the number of sites per category denoted inside the circles. The size of the circles indicates the proportion of neighbouring CTCF sites motifs in tandem, divergent and convergent orientations, respectively. **f** Empirical cumulative density function plots (*left*) for the distance from cohesin-associated CTCF site and their nearest cohesin-associated CTCF counterparts, separated by the motif combination (**e**). The vertical dashed grey line indicates the 5kb minimum distance threshold used for this analysis. The box plots (*right*) show the distribution of these distances for each motif combination (**e**).

The majority of pairs were between two *musculus*-common sites (80%). BL6-specific; however, differed from those expected frequencies, appearing to be nearer to other BL6-specific sites, regardless of their tissue-specificity, significantly more often than expected. This was markedly so in BL6 tissue-shared sites in which 6 times as many neighbouring sites were also BL6 tissue-shared (6.47% vs 1.04% expected) based on their fraction of the total number of CTCF sites (Figure 4.7d). These proportions did not change when the threshold of 5kb minimum distance was relaxed, remaining

almost identical between these two classes. The proportions of motif combinations between neighbouring CTCF binding sites, regardless of evolutionary/tissue-specificity or *cis*/*trans* regulatory variation in the binding sequence, appear to be global features of CTCF binding, and could indicate a deeper evolutionary mechanism driving this pattern of motif orientation they take in short- or long-range.

Like CTCF sites characterized by *cis*/*trans* variation, motif combinations between neighbouring sites were in tandem orientation in 51% of all pairs of CTCF sites regardless of their evolutionary origin or tissue-binding pattern (Figure 4.7e *top*). Pairs of motifs in the convergent orientation were found to be the least common (22%). These differences were found to be statistically significant (χ^2 test, p-value < 0.0001). This disparity in motif orientation, on the other hand, varied among pairs based on the evolutionary origin of their sites (Figure 4.7e *bottom*). Whilst the overall pattern held true for pairs of *musculus*-common sites, pair containing one or two BL6-specific CTCF binding sites were found to either equally be in convergent and divergent orientation (in mixed pairs) or even have more motif combinations in the favourable convergent orientation (in BL6-specific only pairs). Notably, the proportion of tandem motif pairs in those BL6-specific only motif pairs is much higher (55%). Further analysis of the motif combinations of these pairs of cohesin-CTCF regions revealed that, similar to the pattern observed in *cis*/*trans* CTCF sites, tandem/divergent motif orientations are associated with closer pairs of CTCF sites (Figure 4.7f). Motifs in convergent orientations, on the contrary, were further apart (Man Whitney U tests, all p-values < 0.00001).

Taken together, the combination of convergent neighbouring CTCF pairs of sites being further away from each other and occurring only in 22% of all possible motif combinations suggest that chromatin loop formation tend to form between non-neighbouring sites at farther distances. These observations were consistent for all types of CTCF sites, regardless of their evolutionary age, tissue-specificity or the regulatory variation influencing its binding.

4.4 Discussion

4.4.1 Repeat content in *cis*/*trans*-influenced CTCF sites

The CTCF canonical binding motif is known to be carried over by transposable elements in many mammalian lineages[384, 409]. Comparative genomics approaches have shown that lineage-specific CTCF occupancy in mammals is linked to species-specific tRNA-derived SINE expansion. SINE B2 elements have expanded the CTCF binding sites into hundred (in canines) to thousands (in rodents) of new loci in many

mammalian lineages[238, 269, 643]. However, all previous work has been done at large evolutionary distance between the species involved. For example, Schmidt et al.'s work reported waves of SINE-driven expansion of the CTCF binding sites between species separated by 80 million years of divergence time (from canines to rodents and primates)[269].

In this thesis, we investigated this phenomenon in two subspecies that diverged 1 million years ago. This provided us with a unique opportunity to study CTCF evolution by SINE-driven expansion in short evolutionary time. Our analysis of subspecies-specific binding of CTCF in two closely related mice subspecies have confirmed previous findings, but more importantly, have further provided evidence of evolutionary-recent SINE B2-B4 expansion activity in the murine lineage that may well still be ongoing. Results from Chapter 2 have shown that subspecies-specific TE-derived CTCF binding is almost exclusively associated with SINE B2-B4 elements. This is strikingly different from tissue-specific TE-derived TF (CEBPA, FOXA1 and HNF4A) binding in which various superfamilies have played part in contributing to either conserved or divergent TF binding. Though it has been reported that particular TE superfamilies/families drive for the binding of distinct TFs[387], we show here that this process has at least been recently active even for tissue-specific TFs in these two subspecies. TE-derived tissue-specific TF occupancy was observed in a wide range of TE superfamilies, both in subspecies-specific sites and those characterised by informative SNVs that exhibit binding variation between the two subspecies.

Nevertheless, the strong enrichment of SINE B2-B4 elements in CTCF subspecies-specific binding sites show equal contributions of TE superfamilies (particularly SINEs) to that of *musculus*-common binding sites despite significant differences in fraction of binding sites within TEs in general. This supports the idea that the SINE B2-B4 enrichment we observed in evolutionary-young CTCF sites is not the result of an acceleration of the process of TE expansion since the divergence of BL6 and CAST, but rather a continuation of that rodent-specific process that predates the split in their lineage. It is worth noting that any ancient expansion might be undetectable by the methods available to us. Furthermore, some of the *musculus*-common sites are not only conserved between the two subspecies, but given the body of knowledge on CTCF evolution, could also include sites conserved even deeper throughout mammalian lineages. Therefore, the SINE repeat content of these *musculus*-common sites may have been underestimated as any SINE-derived elements in the binding sequence may have diverged too much to be detected by repeat masking algorithms.

Further evidence to recent expansion is provided by analysis of relative age of SINE repeats in both *musculus*-common and BL6-specific binding sites. SINE B2-B4

repeats in BL6-specific sites are much more recent than both *musculus*-common sites and random genomic background, whereas SINEs in *musculus*-common sites belong to an expansion event that had taken place much earlier based on their higher sequence divergence. Furthermore, most of these SINE-derived BL6-specific sites sequences are almost completely covered by SINE repeat elements. Complete masking by SINE repeats could also be a by-product of the small size of BL-specific binding sites, particularly those that are tissue-specific (Figure 4.5f).

On the other hand, CTCF sites characterised by the presence of *cis*/*trans* regulatory variants in BL6, despite harbouring distinctive SNVs to their CAST counterparts, show a repeat enrichment pattern that most closely resembles that of *musculus*-common CTCF sites. Whilst the overall proportion of various TE superfamilies in their sequences is similar to the one discussed above, the fraction of binding sites that are enriched for repeat elements is lower than in non-*cis*/*trans* CTCF sites. SINE B2-B4 elements are again the most predominant type of repeats, but their contribution to the sequence of binding site is 10% lower than in non-*cis*/*trans* sites. The reduction in repeat content of *cis*/*trans* CTCF sites, compared to SNV-free sites, was observed in all superfamilies of TEs, and for all regulatory categories of *cis*/*trans* variants equally.

This was significantly different to liver-specific TFs. First, the most represented TE superfamily in their binding sites were also SINEs (except for CEBPA), exhibiting a marked two-fold enrichment to the genomic background (Figure 4.1a). They were closely followed by LTRs. This is not unexpected, as explained earlier, as TE-derived TF occupancy is driven by different TEs in TF-specific manner. Unlike CTCF, however, *cis*/*trans*-influenced binding of liver-specific TFs was generally associated with higher enrichment in repeat elements than their non-*cis*/*trans* counterparts (except for FOXA1). Nonetheless, the proportions of the different TE superfamilies remained roughly the same between *cis*/*trans* and non-*cis*/*trans* sites across all TFs.

The reduction in repeat content enrichment in *cis*/*trans* CTCF sites in comparison to SNV-free sites could be the result of BL6-specific sites in the latter set, contributing their previously-established (Chapter 2) higher repeat content. Their higher enrichment for TEs in comparison to their non-*cis*/*trans* counterparts shows the opposite pattern, and could indicate an underlying mechanism by which TEs have driven the particular occupancy of these TFs between BL6 and CAST. The binding of these TFs is far less conserved between these subspecies than CTCF[394, 621]. However, *cis*/*trans* sites for these TFs are, by definition, conserved in both orthology (alignability of the sequence between the two subspecies) and occupancy (conserved binding) (only 3-6% of sites show lineage-specificity, see Chapter 3). A possible explanation could be that these TE-derived *cis*/*trans* sites were the results of older TE

activity that had created these sites, which subsequently underwent sequence divergence and acquired SNVs. It is worth mentioning that regardless of the differences described above for CTCF and the liver-specific TFs, the repeat content across all *cis*/*trans* regulatory categories was similar and corresponded to their overall proportions of *cis*/*trans* sites.

Like *musculus*-common CTCF binding sites, *cis*/*trans*-influenced TE-derived CTCF binding was observed to appear in SINE elements belonging to an older cluster of TE expansions. One possibility is that newer SINEs are less likely to have distinguishing SNVs due to the shorter evolutionary time available for them to accumulate over the sequence of SINE elements. Another explanation could be that non-SINE CTCF sites are more likely to have SNVs as their SINE-containing counterparts are resistant to mutation due to their ongoing transposon activity. A recent report proposed that TEs could take advantage of CTCF loop-forming activity to propagate using the host's transcriptional machinery and integrate into genome[644]. Another suggested mechanism involves the association of CTCF and double strand breaks repair to allow TEs to insert themselves in the genome[645]. Perhaps introducing sequence variants represses SINEs ability to transpose effectively, leading to the absence of *cis*/*trans* variants in recent active SINEs.

Prominent TE enrichment in CTCF sites, particularly in subspecies-specific sites, along with the well-established role of CTCF in chromatin loop anchoring is a potential major mechanism for introducing novel higher order chromatin structures. The introduction of CTCF binding sites to new loci via the action of TEs have the potential to divert CTCF from more conserved binding sites, leading to the formation of novel loop contacts. There is evidence that CTCF binding divergence strongly contributes to such a mechanism and SINE-derived CTCF sites have been observed to concentrate at chromatin loops anchors in mouse[219, 238].

In sum, CTCF occupancy evolution in short evolutionary time appear to be considerably attributed to SINE-derived TE insertions that transfer the CTCF binding motif into novel loci. This wave of CTCF binding expansion evidently occurred after the divergence of the two subspecies and may yet still be ongoing. Shared CTCF binding, though enriched for SINE repeat elements compared to liver-specific TFs or genomic background, appear to be the result of a more ancient expansion. This held true for both sets of shared CTCF binding sites whether in *musculus*-common sites, or the subset of shared sites with SNV-derived *cis*/*trans* variation in binding.

4.4.2 CTCF occupancy at active regulatory elements

Transcriptionally active regions in the genome are nucleosome-free, hence are highly accessible when the cells are in interphase[646]. There is accumulating evidence that these regions are enriched for intra-TAD CTCF binding, particularly in enhancer regions [328, 647]. Activating transcription alone has been shown not to be sufficient to result in chromatin insulation, and promoter-enhancer interactions and specific TFs binding, or preceding chromatin conformation changes were suggested to likely contribute to enabling gene expression at specific loci and creating insulation[220]. Recent results proposed the coupling of cohesin loading to lineage-appropriate enhancers and specific genomic locations of CTCF binding to contribute to cell-type-specific control of genome topology[648].

Our analysis showed that CTCF exhibits a preference to bind in intergenic and intronic regions rather than protein-coding sequences or their upstream regulatory elements. This binding pattern is similar across all aspects of CTCF binding investigated in this chapter. Evolutionary young BL6-specific sites had a similar pattern of genome-wide localisation within gene features as *musculus*-common sites. Additionally, *cis/trans* regulatory variation in CTCF occupancy did not produce any discernible differences across any of their categories in this regard.

All BL6 CTCF binding sites, regardless of their evolutionary class, were found to be bound near the same distance to TSSs and downstream genes, mostly just upstream. This is supported by a recent study demonstrating that CTCF sites cluster are significantly closer to TSSs[575]. Distant (in relation to TSS) CTCF binding tends to be more common far upstream than downstream of gene bodies. This could be reconciled with the finding that an RNA strand can bind to and gather multiple CTCF protein molecules near cognate binding sites (the observed clusters of CTCF near TSSs), improves the search for targeted binding sites and enhances the chances of CTCF binding to its sites[649, 650].

These similarities in binding patterns in gene features extend to their co-localisation with histone modifications predictive of active regulatory activity in the liver. CTCF does not appear to co-bind active regulatory regions as much as liver-specific TFs. This makes sense in light of the known biology of CTCF and these liver-specific TFs. CTCF is a genome-wide master regulator that is involved in a wide array of genomic functions, whereas CEBPA, FOXA1 and HNF4A are representative TFs whose evolution and roles are similar to other tissue-specific regulators in mammals that tend to exert their functions by binding to divergent regulatory regions[394, 558]. There is, however, marginal enrichment of CTCF binding in promoter sequences that is significantly higher than would be expected by random. This is consistent with the

observation of CTCF binding just upstream of TSSs. Again, BL6 CTCF binding sites, regardless of their evolutionary class or *cis*/*trans*-acting variants, were bound in enhancer/promoter sequences in equivalent proportions reflecting their contribution to the total number of CTCF in each set. Taken together, these results illustrate that the process of formation of subspecies-specific CTCF binding sites resulted in the maintenance of the same genomic profile and functional signatures as those more conserved (*musculus*-common) or under binding regulatory variation (*cis*/*trans*).

Promoters were found to be the closest active regulatory regions to CTCF sites, both varying in their evolutionary/tissue-specificity type or under *cis*/*trans* regulatory variation, more often than enhancers. However, when an enhancer was found near a CTCF binding site, regardless of classification, the distance between them was always significantly shorter. There were more active regulatory regions that were designated as enhancers than promoters. This could be explained by their tendency to occur closer to TSSs, and their associated promoter sequences. Even though promoters are generally lower in number compared to enhancers, the latter lie further away, possibly resulting in fewer of them associating with CTCF. A previous study showed that the CTCF occupancy at/close to promoters coupled with their nearby cohesin-bound enhancers is associated with upregulation of gene expression[362]. The distance from CTCF sites to their nearest regulatory elements, however, was similar, and there were no significant differences between sites based on their evolutionary/tissue-specificity class, or the category of *cis*/*trans* regulatory variation.

A recent study reported a positive correlation between the enhancer activity as indicated by H3K27ac and CTCF binding, suggesting that CTCF binding influences enhancer activity[627]. They additionally observed that active promoters show higher interaction with CTCF occupancy nearby, and the enhancers interacting with these CTCF sites in turn show higher interaction with active promoters.

Taken together, our findings point to a potential role of evolutionary dynamic CTCF binding, both BL6-specific and *cis*/*trans*-influenced, in facilitating regulatory interaction between distal enhancers and the promoters they regulate. This could potentially explain the reported findings of subtle effects on gene expression observed following targeted CTCF and/or cohesin degradation[651].

4.4.3 CTCF occupancy at TAD-boundary analysis

TADs are chromatin domains that insulate regulatory modules from undesirable interactions, and have been reported to be retained across cell-types and deeply conserved across mammalian lineages[235, 262, 312]. TAD boundary conservation implies that species-specific TE-driven re-wiring is not common. It has

been recently reported that their evolutionary conservation could be the result of clusters of SINE-derived CTCF binding sites at TAD boundaries that help maintain genomic organisation[652]. Mutations affecting TAD loop anchors result in gene expression dysregulation and have been linked to cancer [624, 653]. CTCF is enriched around the anchor loops at TAD boundaries and play a crucial role in the formation and maintenance of TADs.

Our analysis of CTCF occupancy within TADs in sites differing in their evolutionary/tissue-specificity and those affected by *cis*/*trans* regulatory variation has shed light on how short evolutionary time has shaped these sites tendency to associate with TAD boundaries. *Cis*/*trans*-influenced CTCF sites exhibited a strong association with TAD-boundaries. TAD-boundary association for this set of CTCF sites was significantly stronger than general SNV-free CTCF sites. The higher conservation of their binding around TAD boundaries is possibly a product of their sharedness across the two subspecies. Even though they harbour informative SNVs, these sites are, by virtue of their definition, orthologous to their CAST counterparts. Non-*cis*/*trans* CTCF sites, on the other hand, are likely to include sites that are subspecies-specific, TE-derived and otherwise, which in turn may affect their association with TAD boundaries. All categories of *cis*/*trans*-acting variation in CTCF binding sites were found to be equally TAD-boundary associated, showing that TAD-boundary association is independent of allele-specific effects.

This is supported by the observations from CTCF binding around TAD boundaries in *musculus*-common and BL6-specific sites. Whilst the *musculus*-common sites exhibit the characteristic enrichment in the immediate vicinity of TAD boundaries, BL6-specific sites are less TAD-boundary associated, regardless of tissue-specificity. Nonetheless, subspecies-specific still exhibit consistent binding around TAD boundaries, even if their contribution to TAD-boundary associated CTCF sites is proportionally lower than expected based on their overall numbers. These results suggest that even though TAD-boundaries seem to favour more conserved CTCF binding, CTCF sites with dynamic regulatory variation brought on by the introduction of SNVs, or evolutionary young subspecies-specific are not excluded. In fact, sites with *cis*/*trans*-acting variants appear to be more highly associated with TAD-boundaries than general CTCF sites. Recent findings showed that CTCF binding at TAD boundaries are under stronger constraints on their sequence and functional compared to other CTCF sites, independent of their evolutionary conservation between murine lineages. This dynamic evolutionary process was found to be the result of insertions of new species-specific CTCF binding sites in proximity to older ones[575]. Our results provide evidence that these new sites are the result of evolutionary young TE insertions that associate with TAD boundaries in subspecies-specific fashion and, particularly in

the tissue-shared subset of BL6-specific sites, actively associate with cohesin-complex subunits.

It is worth mentioning that despite observed enrichment of CTCF binding near TAD boundaries, the vast majority of CTCF binding sites are widespread in the genome and are located further within TADs structure. The precise nature of the difference between TAD/intra-TAD loop-anchoring sites and the greater number of non-anchoring CTCF sites is still an open question, and role they play in nuclear organisation is still unclear[170, 328, 654]. Given CTCF role to interact with other TFs, bind RNA, and regulate splicing mechanics, these present potential other functions that are not TAD-boundary associated[311, 351, 655].

Our results also indicate that tissue-specific TF binding appears to cluster within certain TADs, whereas ubiquitously bound CTCF genomic distribution within TADs in more uniform. We found out that *cis/trans*-influenced liver-specific TF binding sites (CEBPA, FOXA1 and HNF4A) tend to be found in higher number within TADs than would be expected if randomly distributed. This was apparent in bigger TADs where their numbers considerably increase, in contrast to smaller TADs where the number of observed binding events lower than expected. On the contrary, *cis/trans* CTCF sites were found in higher frequency in smaller TADs, but as the size of TADs increased, their numbers became indistinguishable from the expected frequencies.

This may be explained by their differing mode of genomic action. Tissue-specific TFs bind to regulatory element in genomic regions where they can carry out their function in activating expression of their target genes. These regions are not found genome-wide, instead these are found in defined TADs. These TADs contain genes that tend to be co-regulated as well as genes whose expression is tissue-restricted that are likely to reside within the same TAD. Thus, their binding tends to cluster where their binding is functionally important, resulting in higher occupancy than expected based on random genomic distributions in those TADs. CTCF binding sites, on the other hand, are universally found within almost all TADs where they form the anchors required for TAD and intra-TAD loop formation and maintenance. Their higher number in smaller TADs could be a by-product of the clustering near TAD boundaries discussed above. In large TADs, TAD-boundaries are sufficiently enriched with CTCF binding sites, and the remainder of sites distribute randomly within TADs, where they could be involved in the formation of intra-TAD loops, regulating gene expression, binding RNA or take up other functional roles.

Together with results of repeat content enrichment in CTCF sites, our results support the model for dynamic evolutionary conservation around TAD boundaries, whereby CTCF binding sites associated with TAD-boundaries are more conserved, but

with a substantial contribution of both sites under regulatory variation characterised by the presence of *cis/trans* variants and evolutionary young CTCF sites. The several facet of CTCF occupancy evolution appear to work in tandem to provide stability to functional higher-order chromatin structures. In addition, the observation that, despite significant TAD-boundary association, the majority of CTCF sites reside well within the TADs yet still strongly associate with cohesin-complex subunits suggest potential involvement in the formation of intra-TAD loop anchors. This, coupled with proximity to regulatory active regions, makes them candidate for the formation of novel loops that mediate promoter-enhancer contacts and gene expression. Indeed, a recent report indicates that TE-derived loop anchors are associated with variable chromatin looping across species and cells[644].

4.4.4 CTCF recruitment of cohesin-complex proteins

To carry out its function in establishing loop contacts and setting up insulated regulatory domains, CTCF interacts with cohesin-complex proteins found near the TAD boundaries[235, 328, 625, 656].

Our analysis revealed that CTCF sites that are under *cis/trans*-acting variation strongly co-localise with cohesin-complex subunits. Liver-specific *cis/trans* TFs, on the other hand, show a variable degree of co-localisation with cohesin. Their pattern of cohesin co-localisation, however, showed that they mostly associate with cohesin-non-CTCF sites (CNC). CNC sites are often at tissue-specific promoters and enhancers[224]. These same liver-specific TFs have previously been found to associate with CNC sites to help stabilize large TF complexes by facilitating the binding of master regulators and enhancer-markers, resulting in the upregulation liver-specific gene expression[362]. CNC sites bound by TF binding sites facilitate distinct gene expression states by promoting cohesin-driven interactions within an existing chromatin structures[657]. These results, combined with previous observation of regulatory elements enrichment (particularly in enhancer regions) and their tendency to cluster within certain bigger TADs, suggest that liver-specific binding of these TFs reflects significant hallmarks of functional activity even when under *cis/trans* regulatory variation. These sites have also been shown (see Figure 4.1b) to have a higher enrichment in repeat content in their binding sites compared to their SNV-free counterparts, potentially implicating TE-derived TF occupancy in promoting these functional signatures in cell-type specific context.

Our analysis has also shown that *cis/trans* CTCF co-binding with cohesin is predominantly TAD-boundary associated, regardless of the type of *cis/trans* regulatory variation characterising the binding site. In terms of their evolutionary/tissue-specificity class, however, we observed significant difference in co-

localisation with CTCF. Whereas *musculus*-common sites were found to be strongly associated with cohesin-complex subunits, BL6-specific sites differed based on their tissue-specificity. BL6 tissue-specific sites exhibited striking reduction in co-localisation with cohesin-subunits, whereas BL6 tissue-shared sites showed a much stronger association with higher recruitment of cohesin.

Furthermore, stronger CTCF sites were found to co-localise with cohesin more often than weaker ones regardless of evolutionary/tissue-specificity class. However, due to the majority of BL6 tissue-specific sites exhibiting generally lower ChIP signal, fewer of them are associated with cohesin in total, explaining the apparent reduction in cohesin co-localisation in this subset of CTCF binding sites. BL6 tissue-shared site, on the other hand, exhibit a uniform distribution in ChIP-signal with roughly equal proportion at each tier of binding signal, hence they have stronger signal and cohesin recruitment.

Taken together, our findings from the analysis of CTCF co-localisation with cohesin in evolutionary/tissue-specificity and *cis*/*trans* variation contexts provide some clues as to what factors are involved in this process. First, the more conserved a CTCF site is, the higher the co-localisation with cohesin subunit. Orthologous CTCF binding sites under *cis*/*trans* regulatory variation were found to show a similar pattern of association with cohesin as *musculus*-common sites, despite the presence of sequence variants by virtue of their subspecies-sharedness (only 1.7% of sites showed lineag-specific binding. See Chapter 3). Even though BL6-specific sites showed reduced association with cohesin, as a subset of these acquired tissue-wide occupancy, their co-localisation with cohesin rose significantly to almost equal that of a more conserved status. A second factor influencing CTCF co-localisation with cohesin was ChIP signal. When sufficiently high, all CTCF sites were found to equivalently co-localise with cohesin regardless of their conservation state. However, as their signal dwindled, the level of their association with cohesin dropped, and depending on the fraction of sites for each class of CTCF within the tiers of ChIP signal, the overall level of association with cohesin was affected.

A couple of computational approaches have attempted to predict intra-TAD loop formation based on CTCF and cohesin peak strength as the primary predictor[658, 659], lending support to the idea of CTCF signal as a proxy to its association with cohesin and its involvement in chromatin remodelling.

A study estimated that ~90% of DNA loops are associated with CTCF and cohesin binding, in which 92% comprised of CTCF anchors whose motifs were in convergent orientation[219]. Convergent orientation has been a consistent feature for identifying chromatin loops using computational methods[285, 365]. However, loops

exhibiting other orientations have been observed to represent as few as 8% (in loop domains), and as many as 20% (in insulated Neighbourhoods)[219, 624].

Our investigation into the motif combinations of neighbouring CTCF sites showed equivalent results. In CTCF-and-cohesin (CAC) binding sites under influence of *cis/trans*-acting variants, half of *cis/trans* CAC sites are in tandem orientation, which is expected based on the frequencies of the +/- motif proportions. However, almost a third are in divergent orientation, and fewer than expected are in the preferable convergent orientation most common in chromatin loops. Pairs of neighbouring CTCF sites in convergent orientation are additionally found further away in their distance than the other two orientations. These findings did not differ across the different regulatory categories of *cis/trans* variants.

The results of the same approach in CAC sites differing in their evolutionary/tissue-specificity classification were generally similar, particularly in the ratios of motif combinations in pairs of neighbouring CTCF sites and their distance to each other. There were, however, some notable differences. BL6-specific CAC sites were more likely to be nearer BL6-specific sites (particularly tissue-shared ones) than expected. Furthermore, even though more pairs of motifs were in divergent than convergent orientation, this was only limited to all *musculus*-common pairs. Mixed or BL6-specific pairs either exhibited similar proportion of the two (tissue-specific) or appeared more convergent orientation (tissue-shared).

These results indicate that CTCF sites, in association with cohesin, do not generally exhibit favourable motif orientation in short distances between pair of neighbouring sites. Given that median size of shorter loops in complex nested intra-TADs is 185 kb[219], it is most likely that these loops take place between CTCF-cohesin anchors separated by several intervening CTCF motifs not in the preferable orientation. This could also be seen in the increased distance separating pairs of motifs that do occur in the convergent orientation in both sets of CTCF sites investigated. The fact that motif orientations between pairs of neighbouring CTCF sites did not differ based on their evolutionary/tissue-specificity, or the type of *cis/trans* variant present is a testament to the plasticity of CTCF sites undergoing dynamic evolutionary rewiring to maintain existing structures and functions, as well as readily adopting them when new sites arise. This provides a rich source for lineage- and cell-type specific genomic innovation.

Chapter 5

Conclusions and future directions

In this thesis, I have explored the regulatory evolution of a master genome regulator, the CTCF protein, in two mouse subspecies separated by a short evolutionary divergence time, which participate in transcriptional regulation of gene expression. The analyses described in this thesis were conducted using computational approaches on a genome-wide scale to study the binding pattern of CTCF and the biological implications of such occupancy in a tissue- and subpecies-specific manner. These projects aimed to understand how evolutionary divergent CTCF binding sites are shaped by evolutionary forces and what patterns their occupancy exhibits in both subpecies-specific and F1 hybrid biological contexts. Combining computational methods with advancement in next-generation sequencing (NGS) techniques, I was additionally able to leverage the availability of high-quality genome sequences, and the publicly available data, either from large-scale consortia efforts, or from data repositories of published work by our group and others.

The project covered in Chapter 2 investigated the mechanisms and possible functional consequences underlying the birth of evolutionary-young CTCF binding sites in short evolutionary time between two mouse subspecies in the murine lineage. A main finding from this investigation is that the evolution of novel CTCF binding sites is indeed driven, at least in half the cases, by transposable elements action, particularly from the B2-B4 family of SINE elements. In addition to confirming earlier reports[269, 384], this study elucidated that the process continued after the divergence of the two species, leading to the proliferation of thousands of new CTCF binding sites. Another key result concerned a subset of those young subpecies-specific sites which displayed consistent binding across multiple tissues in *M. musculus domesticus*

(BL6), suggesting that they may possess enhanced transcriptional regulatory functionality. In general, all subspecies-specific CTCF sites bore all genomic features and occupancy patterns of conserved CTCF sites, suggesting that they may all be functional. Furthermore, the study uncovered a novel BL6-specific immune locus consisting of a 15-gene cluster on chromosome 4, formed by a recurrent regulatory architecture consisting of a CTCF binding site and an interferon gene. These results provide evidence on how quickly evolutionary-young CTCF sites start to demonstrate several functional signatures shortly after their integration into the genome.

Chapter 3 focused on the effects of genetic sequence changes on CTCF binding site occupancy variation and the underlying mechanisms causing this variation. Using the F1 hybrid of the same two mouse subspecies in the first project to investigate CTCF binding divergence in response to *cis*- and *trans*-acting variation. The findings of this study demonstrate pervasive *trans* variation on CTCF occupancy influencing its allelic-specific binding in hybrid contexts in addition to established *cis* effects. The inheritance of these *cis* and *trans* effects, though additive, is visibly influenced by dominant effects. Lineage-specific binding events are more scarce owing the overwhelming conservation of CTCF binding between the two species demonstrated in the first study. Although CTCF is well-known to be heavily involved in long-range promoter-enhancer interaction via looping (see Chapter 1), CTCF binding sites under *cis*-acting variation do not display any measurable signal coordination over short or long genomic distance with other CTCF sites in an allele-specific manner. Taken together, the results of this project indicate the pervasive influence of *trans*-acting variation which contribute to the complex pattern of regulatory effects imposed on and by the universally expressed CTCF.

To establish the extent and differences in regulatory potential of CTCF occupancy in short evolutionary time, Chapter 4 investigated several functional facets of CTCF binding sites either on the basis of their evolutionary/tissue-specificity (Chapter 2) and the binding variation they exhibit in the form of *cis*/*trans* variants (Chapter 3). Namely, we compared and contrasted the pattern observed between both sets of CTCF sites in terms of their repeat elements content and age, their binding in active regulatory elements, TAD-boundary association and their interaction with cohesin-complex proteins. Our results support our earlier findings of a recent, post-divergence SINE-driven expansion of CTCF binding sites in subspecies-specific manner that is absent even from sites shared between the two strains, yet characterised by SNV insertion. We also further illustrate the dynamic evolution of CTCF association with TAD-boundaries, where conservation of binding coincides with substantial contribution of both sites under *cis*/*trans* regulatory variation and subspecies-specific CTCF sites. This was coupled with strong interactions with cohesin-complex proteins in both *cis*/*trans*-influenced sites and evolutionary young, tissue-shared sites. Taken

together, the functional regulatory aspects of evolutionary dynamic CTCF sites of all types considered in this thesis indicated that whilst the maintenance of pre-existing higher-order chromatin structures is evident, the evolution of CTCF occupancy, whether by SNVs or TE-derived insertions, provides a template for lineage- and tissue-specific genomic innovation.

These results, put together, add to the increasing literature on the regulation of gene expression. *Cis*-regulatory elements such as promoters and enhancers are involved in a multitude of long-range contacts via a looping mechanism facilitated by CTCF and its associated cohesin complex in order to load the transcriptional machinery and allow the physical interaction between distal and proximal regions of the circuitry[660]. While CTCF is crucial to this process of chromatin interaction and 3D architectural organisation during regulation of bilaterian animals gene expression, in non-bilaterian animal, this mechanism of long-range interactions between constituents of the regulatory machinery is lacking CTCF[273]. It has been proposed that in large genomes such as those of mammals, *cis*-regulatory elements can be distantly located to the target genes they regulate. Thus, they contain CTCF-dependent boundaries to allow different combination of loci to display different chromatin/compartment states, and enable long-range interactions while effects from flanking TADs are insulated[323, 367, 661]. This could explain the observed pattern of proliferation of CTCF sites during evolutionary divergence. These evolutionary young sites, with their readily acquisition of typical functional signatures of CTCF, introduce a degree of redundancy in the system that may protect it from the potential effects of genetic and environmental abnormalities[662].

Furthermore, although chromatin looping forming TADs usually appear conserved between tissues, cell-types and even in syntenic regions of related species, it has been recently found that some CTCF-mediated chromatin loops are not constitutive, but rather tissue-specific[663]. The subset of CTCF-mediated loops may have risen due to tissue-specific CTCF binding, mediated by other tissue-specific epigenetic modifications or transcription factors[664]. Alternatively, tissue-wide, constitutive CTCF binding sites may be involved in tissue-specific interactions mediated by the presence of additional looping co-factors to bring about cell-type-specific CTCF-mediated interactions[665]. The subset of species-specific CTCF binding sites identified in the first study (Chapter 2) may provide an evolutionary mechanism for such tissue-specific interactions to originate. Since conserved CTCF sites and TAD domains do not appear to diverge significantly, especially between closely related species, such novel species-specific and tissue-variable sites offer a platform for cell-type-specific interactions to occur.

The analyses conducted in the second project was facilitated by the large number of SNVs (~19 M) previously identified[257], which is comparable to the number of SNVs between human populations[666], between the two mouse subspecies, allowing sufficient resolution to detect measurable difference in CTCF binding intensities. Dissecting the effects of *cis*- and *trans*-acting variant on CTCF binding intensity provides an opportunity to understand the consequences of sequence changes to the binding sites on CTCF function. This is particularly medically relevant. Although somatic mutations in the CTCF coding gene have been detected in cancer[667], a great number of recurrent mutations were identified in CTCF binding sites in a number of human cancers[668]. Some SNPs in the CTCF binding site have also been shown to increase disease susceptibility by impacting methylation levels at differentially methylated CTCF-binding sites[669]. Mutations and SNPs can additionally alter gene expression and tumour progression by disrupting the CTCF-mediated folding of chromatin[647]. A study combining Hi-C and ChIP-seq has shown that, whilst super enhancers (SEs) hub and non-hub enhancers share common histone marks, hub enhancers appear to associate with CTCF and cohesin binding sites and disease specific SNPs[665]. Although CTCF insulates super enhancers from non-target genes by forming boundaries between loci, this insulation action is fluid, and dependent on the strength of CTCF occupancy or "insulation score" at the SE boundary. Such occupancy-dependent insulation score could be investigated by measuring the relative contribution of *cis*- and *trans*-acting variation on candidate regions.

The analyses carried out in this thesis also offer some directions for future work. At the moment, publicly available Hi-C protocols in matched samples do not offer a resolution high enough to capture long-range chromatin interactions involving a pair of CTCF sites to the level of their binding motif. This is critical, particularly in the case of many of the species-specific CTCF binding sites identified in Chapter 2, as they often occurred in clusters with nearby conserved CTCF sites in regions a few kb long. Such regions could not be resolved using available data. This hampered the discovery of such an association of many of these sites with novel chromatin looping. Nevertheless, advances in chromatin capture technologies, super-resolution imaging, and sophisticated *in silico* modelling are likely to soon bridge this gap in knowledge and permit us to revisit this aspect in the future in order to verify whether such sites are indeed an evolutionary venue for regulatory innovation. For example, a recent publication reported the development of a machine learning algorithm, CTCF-MP, that combines functional genomic signals from CTCF ChIP-seq and DNase-seq to make accurate predictions on whether a pair of convergent CTCF binding motifs would form a loop[670]. The algorithm is based on word2vec, a popular word embedding model in natural language processing, and only utilizes sequence-based features to inform the model if a convergent CTCF motif pair have the capacity to for loop formation in a single cell type and also across different tissues.

In addition, the application of ChIP-seq to study protein-DNA interactions requires tissue sample with a large number of cells and relies on several parameters such as the choice of single- vs. paired-end tags and antibody quality. Advances in synthetic biology may allow the development of tailor-made antibodies with better affinity to the protein of choice, reducing the level of biological noise associated with ChIP-seq. A 2015 ChIP-seq protocol was developed to allow profiling of protein-DNA interactions with low input of cells[671].

A powerful novel venue to develop a better understanding of the heterogeneity inherent to the system and cell-to-cell variability is to able to perform single-cell ChIP-seq. A major breakthrough came by in 2017 when Stevens et al.[672] were able to study the 3D structures of complete genomes, the pattern of genome folding, TADs and loops at $<$ 100 kb resolution in single cell level, leveraging data from the 3C method. They also reported evidence that TADs and CTCF cohesin loops form in partial cells and exhibit dynamic structural changes and variations in 12-62% of the cells. Another advancement came from Ren et al.[627] who combined gene editing technology with single-cell flow cytometry and single-molecule RNA-FISH assays to demonstrate that CTCF helps to stabilize enhancer-promoter interactions, resulting in maintenance of minimal variations of gene expression.

Finally, despite the versatile regulatory functions of CTCF as a transcriptional activator/repressor, an insulator/boundary element binding factor, or a regulator of genomic imprinting and higher-order chromatin folding, the molecular mechanisms of CTCF in cell differentiation and disease development is still a work in progress. However, the progress made in our knowledge and understanding of the 3D conformation of the genome has been growing, and the plethora of loop types, shapes, and functions within the chromatin has become more complex. The identification and validation of CTCF-mediated mechanisms that affect biological function and transcriptional regulation in species-, tissue- and cell-specific fashion is a major step in the direction of untangling this complexity.

Publications

The following publication is the result of my doctoral work in this thesis:

- Azazi D, Mudge JM, Odom DT, Flicek P: **Functional signatures of evolutionarily young CTCF binding sites**. *BMC Biology* 2020, **18**: 132. https://doi.org/10.1186/s12915-020-00863-8

Appendix 1

Repeat masking results of TF binding sites

Table 1: CTCF

Evolutionary type	Repeat Element	Class	Family	No. Elements	Bases masked	% of Sequence
musculus-common	Interspersed repeats	SINEs:	Alu/B1	1101	87722	1.00%
			B2-B4	6765	980357	11.12%
			IDs	88	5469	0.06%
			MIRs	142	12010	0.14%
		LINEs:	LINE1	858	95385	1.08%
			LINE2	71	6336	0.07%
			L3/CR1	8	580	0.01%
		LTR	ERVL	232	29157	0.33%
			ERVL-MaLRs	581	77235	0.88%
			ERV_classI	117	17128	0.19%
			ERV_classII	256	33712	0.38%
		DNA	hAT-Charlie	253	25839	0.29%
			TcMar-Tigger	35	4684	0.05%
		Unclassified:		116	22599	0.26%
	Small RNAs:			73	6746	0.08%
	Satellites:			21	2174	0.02%
	Simple repeats:			4216	167652	1.90%
	Low complexity:			508	23979	0.27%
	Total			32155	1605391	18.21%
BL6-specific	Interspersed repeats	SINEs:	Alu/B1	129	10240	0.76%
			B2-B4	3599	455525	33.66%
			IDs	12	692	0.05%
			MIRs	6	651	0.05%
		LINEs:	LINE1	252	29837	2.20%
			LINE2	7	543	0.04%
			L3/CR1	0	0	0.00%
		LTR	ERVL	25	2284	0.17%
			ERVL-MaLRs	147	16705	1.23%
			ERV_classI	117	17898	1.32%
			ERV_classII	251	35358	2.61%
		DNA	hAT-Charlie	28	2616	0.19%
			TcMar-Tigger	3	300	0.02%
		Unclassified:		21	2964	0.22%
	Small RNAs:			16	2373	0.18%
	Satellites:			200	46504	3.44%
	Simple repeats:			554	24255	1.79%
	Low complexity:			35	1737	0.13%
	Total			7021	651800	48.16%
CAST-specific	Interspersed repeats	SINEs:	Alu/B1	67	5145	0.67%
			B2-B4	1842	232983	30.19%
			IDs	4	238	0.03%
			MIRs	5	443	0.06%
		LINEs:	LINE1	209	23946	3.10%
			LINE2	3	188	0.02%

			L3/CR1	2	131	0.02%
		LTR	ERVL	29	3392	0.44%
			ERVL-MaLRs	124	13294	1.72%
			ERV_classI	31	4602	0.60%
			ERV_classII	140	19390	2.51%
		DNA	hAT-Charlie	21	1681	0.22%
			TcMar-Tigger	1	125	0.02%
		Unclassified:		22	2884	0.37%
	Small RNAs:			2	216	0.03%
	Satellites:			23	4481	0.58%
	Simple repeats:			262	9667	1.25%
	Low complexity:			19	799	0.10%
	Total			4730	324033	41.98%

Table 2: CEBPA

Evolutionary type	Repeat Element	Class	Family	No. Elements	Bases masked	% of Sequence
musculus-common	Interspersed repeats	SINEs:	Alu/B1	1505	134460	1.27%
			B2-B4	2040	234401	2.21%
			IDs	238	16216	0.15%
			MIRs	811	86345	0.81%
		LINEs:	LINE1	1341	220361	2.08%
			LINE2	305	30240	0.29%
			L3/CR1	47	5254	0.05%
		LTR	ERVL	450	71587	0.68%
			ERVL-MaLRs	1401	220386	2.08%
			ERV_classI	304	52640	0.50%
			ERV_classII	1007	192893	1.82%
		DNA	hAT-Charlie	531	73582	0.69%
			TcMar-Tigger	84	12851	0.12%
		Unclassified:		90	16667	0.16%
	Small RNAs:			135	11464	0.11%
	Satellites:			12	966	0.01%
	Simple repeats:			2393	82960	0.78%
	Low complexity:			227	9274	0.09%
	Total			39196	1499055	14.14%
BL6-specific	Interspersed repeats	SINEs:	Alu/B1	195	16629	1.16%
			B2-B4	388	43896	3.07%
			IDs	34	2307	0.16%
			MIRs	64	6077	0.43%
		LINEs:	LINE1	242	37972	2.66%
			LINE2	24	2390	0.17%
			L3/CR1	9	943	0.07%
		LTR	ERVL	75	12583	0.88%
			ERVL-MaLRs	273	40280	2.82%
			ERV_classI	112	19000	1.33%
			ERV_classII	500	85564	5.99%
		DNA	hAT-Charlie	56	7208	0.50%
			TcMar-Tigger	4	563	0.04%
		Unclassified:		24	4618	0.32%
	Small RNAs:			44	4545	0.32%
	Satellites:			205	85374	5.97%
	Simple repeats:			295	11121	0.78%
	Low complexity:			18	709	0.05%
	Total			6733	383229	26.81%
CAST-specific	Interspersed repeats	SINEs:	Alu/B1	1570	133639	1.33%
			B2-B4	2383	263074	2.61%
			IDs	237	16362	0.16%
			MIRs	525	51411	0.51%
		LINEs:	LINE1	2944	420441	4.17%
			LINE2	257	26643	0.26%

			L3/CR1	31	3500	0.03%
		LTR	ERVL	496	70792	0.70%
			ERVL-MaLRs	1685	248848	2.47%
			ERV_classI	557	93316	0.93%
			ERV_classII	2083	386285	3.83%
		DNA	hAT-Charlie	540	69936	0.69%
			TcMar-Tigger	86	11871	0.12%
		Unclassified:		113	18509	0.18%
	Small RNAs:			116	10594	0.11%
	Satellites:			1373	519949	5.16%
	Simple repeats:			2292	80701	0.80%
	Low complexity:			277	12353	0.12%
	Total			43546	2458654	24.40%

Table 3: FOXA1

Evolutionary type	Repeat Element	Class	Family	No. Elements	Bases masked	% of Sequence
musculus-common	Interspersed repeats	SINEs:	Alu/B1	4099	386810	1.86%
			B2-B4	4942	601384	2.90%
			IDs	510	34911	0.17%
			MIRs	1216	128461	0.62%
		LINEs:	LINE1	2076	361939	1.74%
			LINE2	696	75589	0.36%
			L3/CR1	83	9548	0.05%
		LTR	ERVL	743	129282	0.62%
			ERVL-MaLRs	2293	389976	1.88%
			ERV_classI	537	111838	0.54%
			ERV_classII	1626	340159	1.64%
		DNA	hAT-Charlie	790	113895	0.55%
			TcMar-Tigger	166	25737	0.12%
		Unclassified:		156	36521	0.18%
	Small RNAs:			240	21490	0.10%
	Satellites:			71	6703	0.03%
	Simple repeats:			7023	271884	1.31%
	Low complexity:			705	31382	0.15%
	Total			57301	3123172	15.04%
BL6-specific	Interspersed repeats	SINEs:	Alu/B1	788	71476	1.83%
			B2-B4	1177	136828	3.50%
			IDs	72	4917	0.13%
			MIRs	140	12938	0.33%
		LINEs:	LINE1	516	84574	2.16%
			LINE2	93	8858	0.23%
			L3/CR1	8	666	0.02%
		LTR	ERVL	160	25675	0.66%
			ERVL-MaLRs	625	95394	2.44%
			ERV_classI	290	71421	1.83%
			ERV_classII	689	129920	3.32%
		DNA	hAT-Charlie	113	15477	0.40%
			TcMar-Tigger	14	1579	0.04%
		Unclassified:		59	13064	0.33%
	Small RNAs:			84	8949	0.23%
	Satellites:			92	24602	0.63%
	Simple repeats:			1340	56197	1.44%
	Low complexity:			122	6664	0.17%
	Total			16097	774402	19.82%
CAST-specific	Interspersed repeats	SINEs:	Alu/B1	1812	166847	1.91%
			B2-B4	2539	296600	3.40%
			IDs	204	13954	0.16%
			MIRs	421	43706	0.50%
		LINEs:	LINE1	2169	346026	3.96%
			LINE2	260	26039	0.30%

			L3/CR1	36	3995	0.05%
		LTR	ERVL	459	70981	0.81%
			ERVL-MaLRs	1543	247238	2.83%
			ERV_classI	621	118649	1.36%
			ERV_classII	1605	314595	3.60%
		DNA	hAT-Charlie	387	52931	0.61%
			TcMar-Tigger	71	10439	0.12%
		Unclassified:		114	22021	0.25%
	Small RNAs:			112	10653	0.12%
	Satellites:			500	97836	1.12%
	Simple repeats:			3158	124788	1.43%
	Low complexity:			327	15687	0.18%
	Total			32625	2000771	22.91%

Appendix 2

The effect of biological replicates availability on *cis/trans* variation in TFs

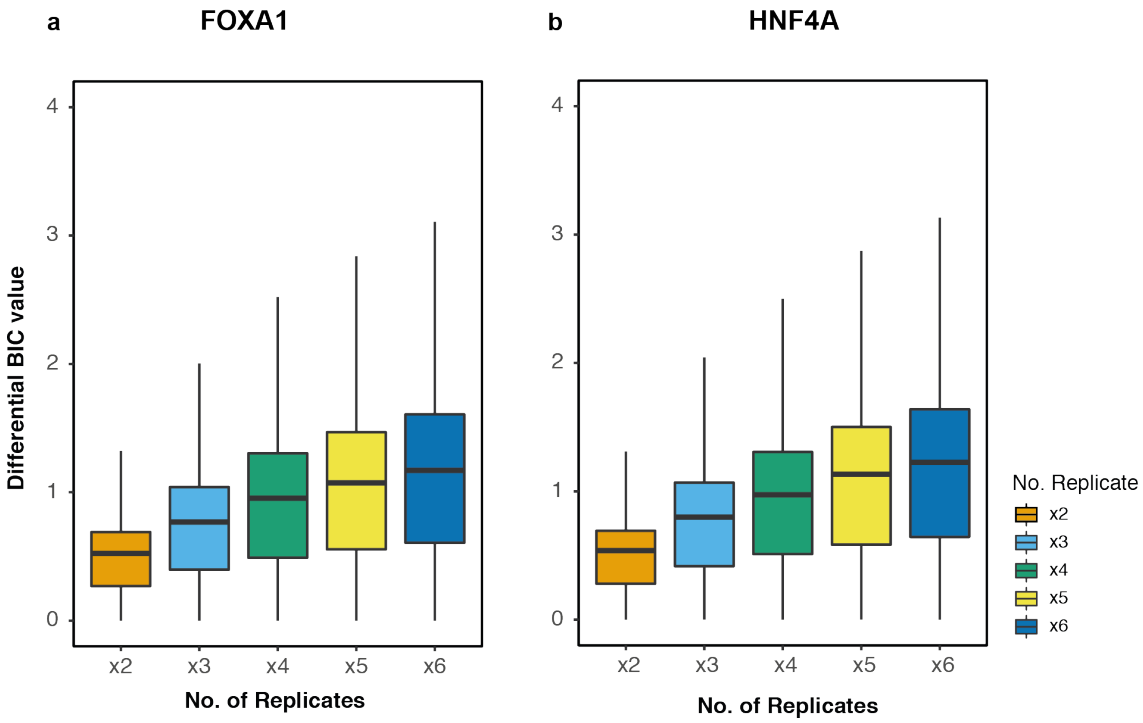


Figure S2.1: Improvement in category-assignment quality with increasing replicate number

Boxplots of the differential BIC values for all TF binding sites under *cis/trans* regulatory variation in ascending number of libraries in FOXA1 (**a**) and HNF4A (**b**).

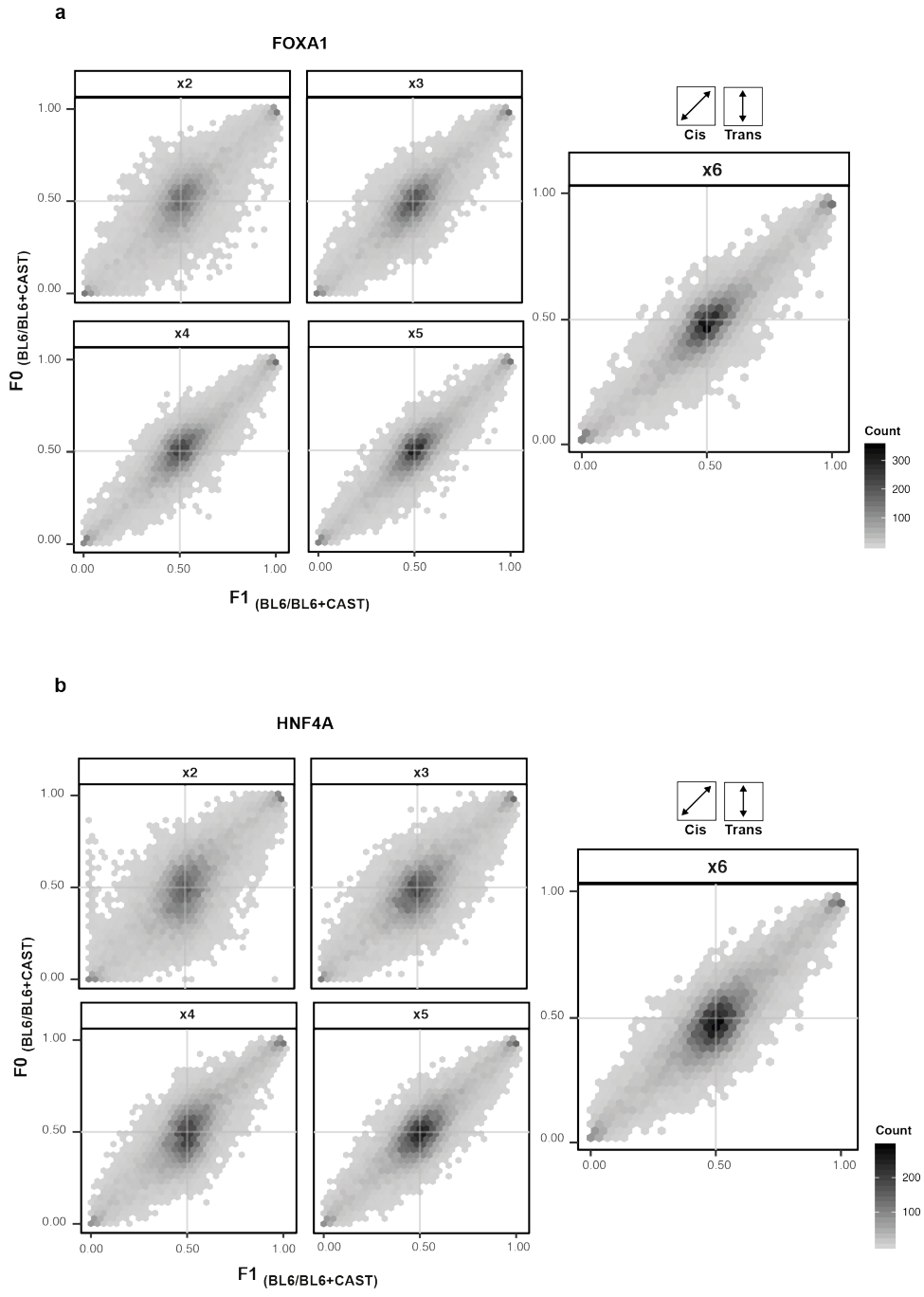


Figure S2.2: Increasing *cis* effect size on binding ratios between F1 and F0

Hexagonal heatmaps for the F0 versus. F1 binding intensity ratios (BL6 vs. CAST) for every *cis*/*trans* site in 2 to 6 biological replicates in FOXA1 (**a**) and HNF4A (**b**).

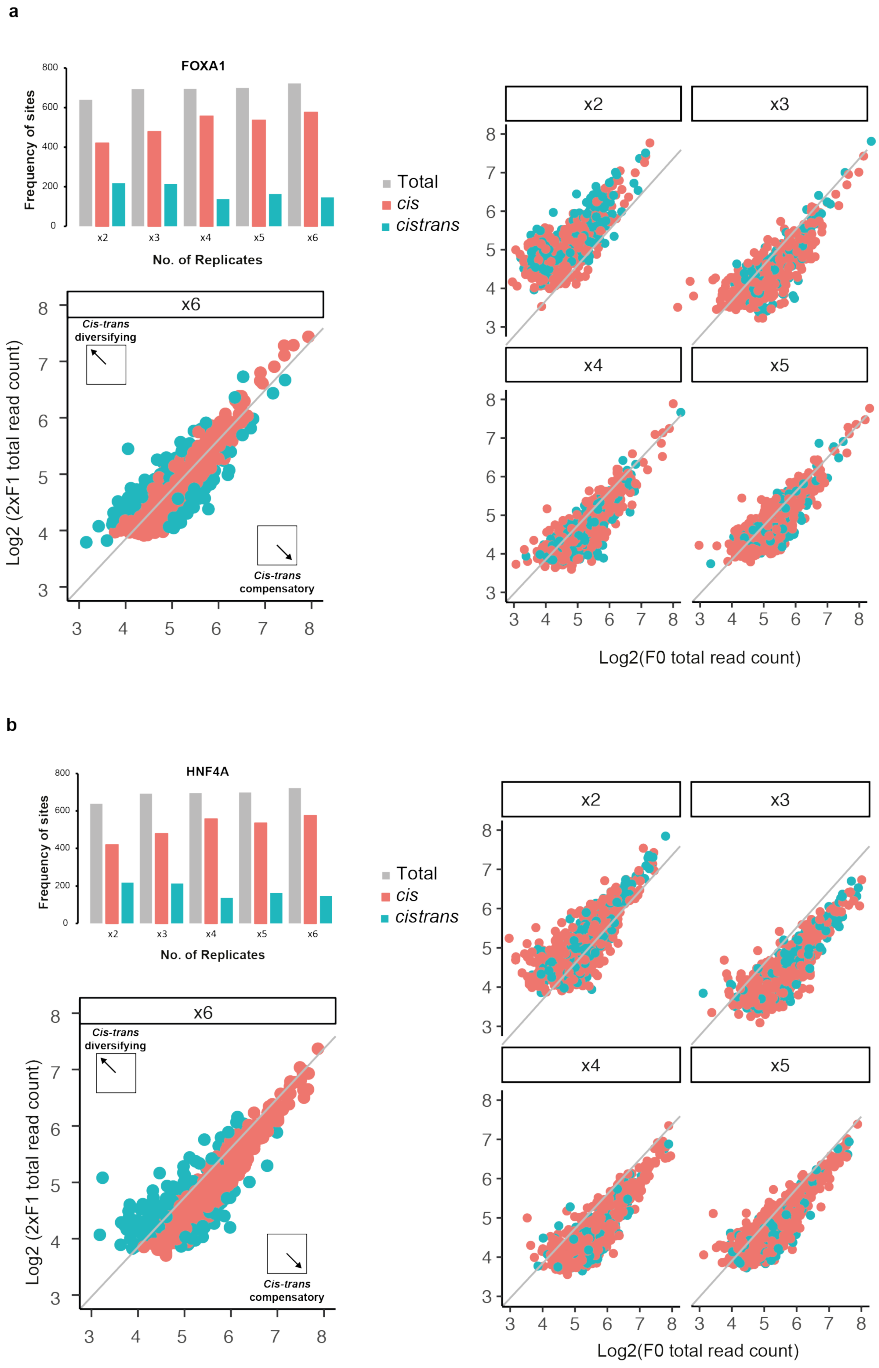


Figure S2.3: Evolutionary direction of lineage- specific sites is enhanced with increasing replicate number

Bar chart of the number of TF binding sites classed as lineage-specific for both *cis* and *cistrans* variants in 2 to 6 biological replicates in FOXA1 (**a**) and HNF4A (**b**), with scatter plots of average log2 F0 total read counts against average log2

F1 read count (BL6 plus CAST allele) multiplied by 2, using averages in 2 to 6 biological replicates.

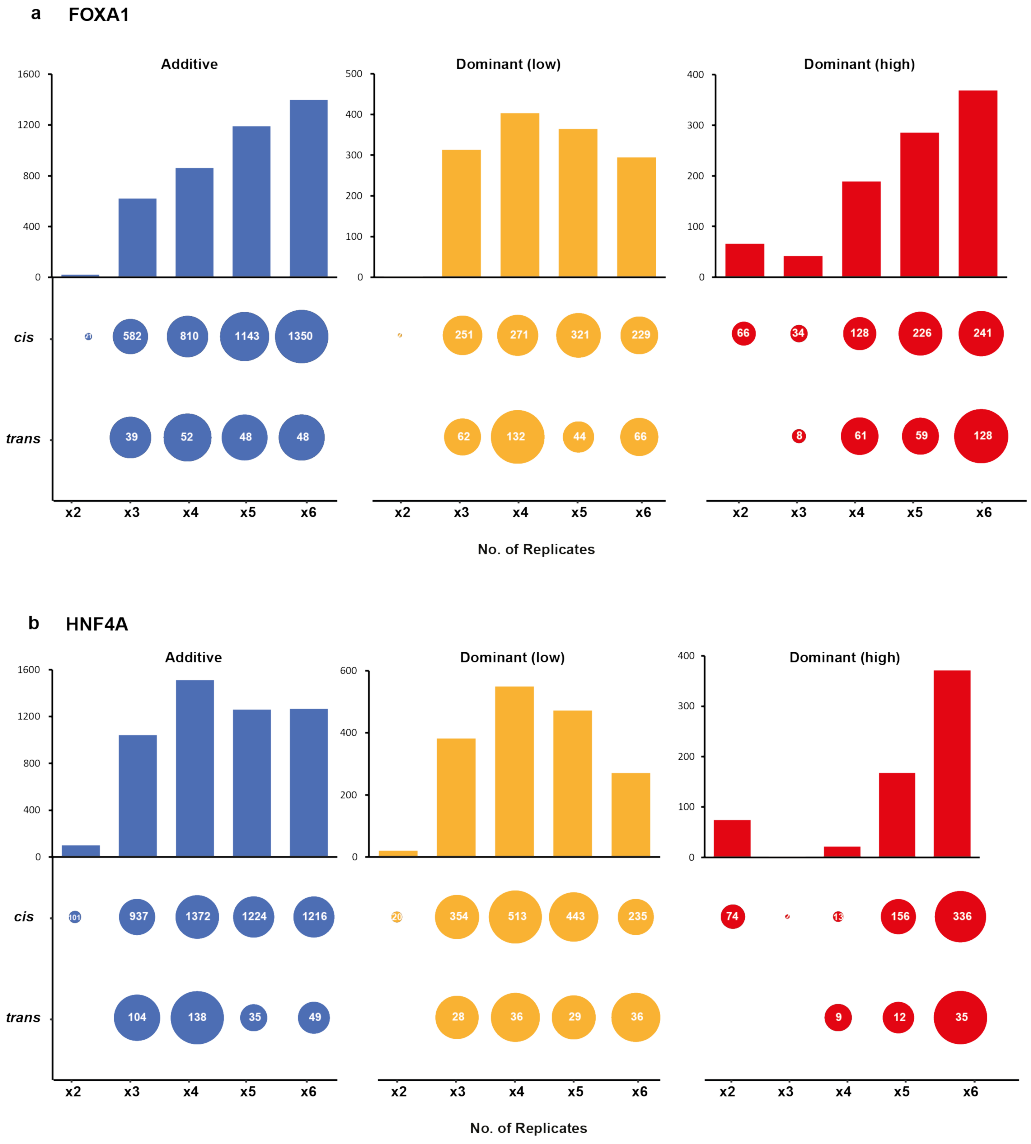


Figure S2.4: Availability of more libraries markedly improves the determination of inheritance patterns of in *cis*/*trans* TF sites

Bar plot of the number of binding sites based on their assigned mode of inheritance in ascending number of biological replicates in FOXA1 (**a**) and HNF4A (**b**). The circles illustrate the make-up of each mode of inheritance by the type of *cis*/*trans* variation acting on the binding site for each number of replicates, with the number of sites per category denoted inside the circles.

Appendix 3

TE-masking and genomic features of *cis*/*trans* TF binding

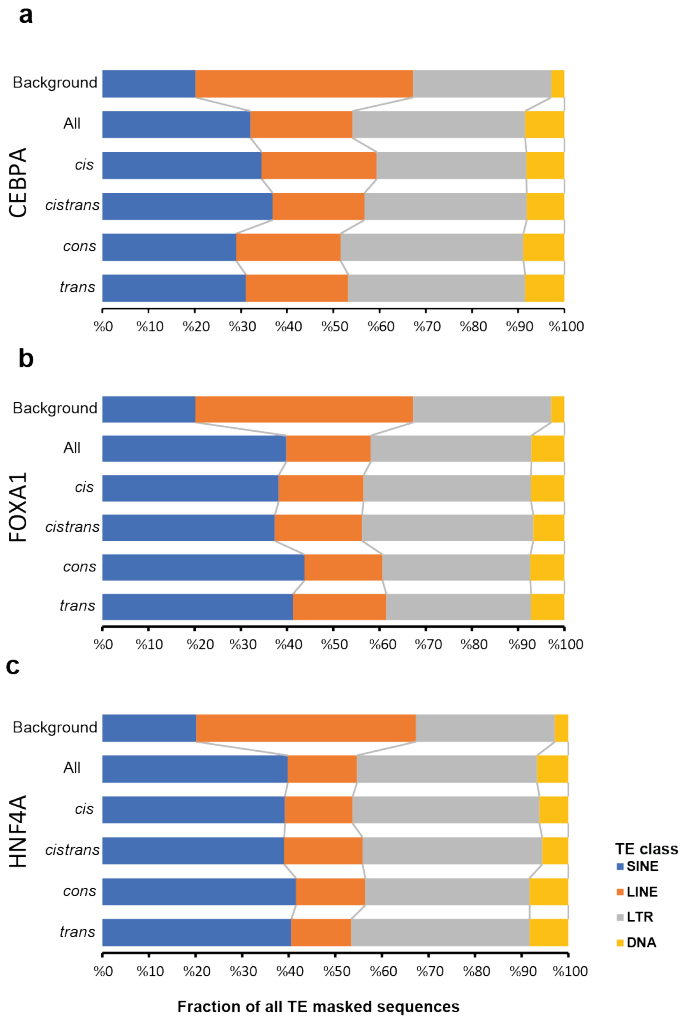


Figure S3.1: TE-derived binding in *cis*/*trans*-influenced binding of liver-specific TFs

Top horizontal bar chart shows the fractions of major TE classes in *cis/trans* TFs binding sites that are masked by repeat elements in CEBPA (**a**), FOXA1 (**b**) and HNF4A (**c**). The top bar refers to the percentage each TE class occupies in all repeat masked sequences in the BL6 mouse genome as a background.

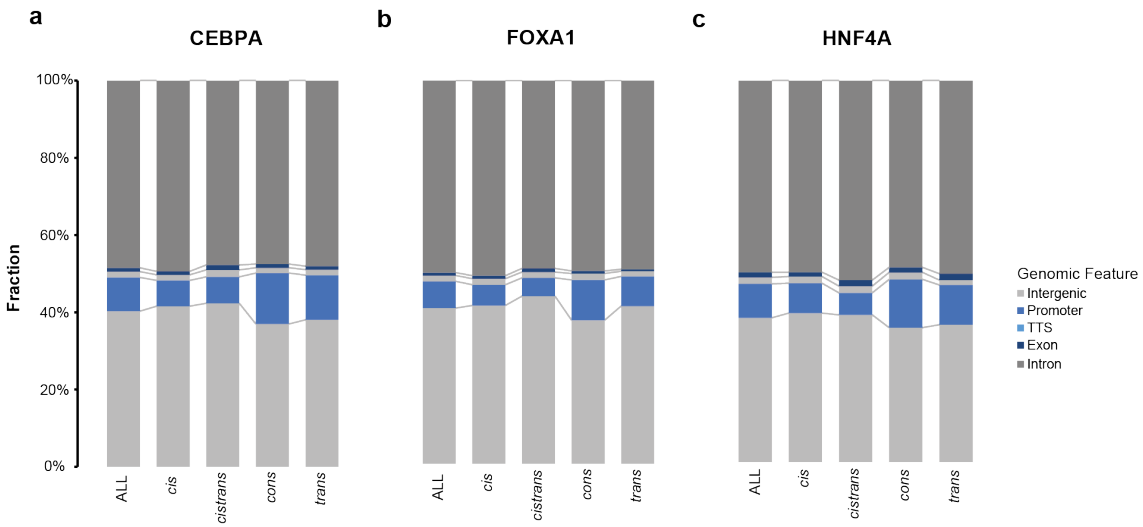


Figure S3.2: Enrichment of *cis/trans* CTCF sites at proximal active regulatory elements suggest potential regulatory activity

The bar charts shows the fraction at which different TFs are found in the respective genomic element in CEBPA (**a**), FOXA1 (**b**) and HNF4A (**c**), with a breakdown by the type of *cis/trans* variant present at the binding site.

Appendix 4

Scripts & Pipelines

Alignment and Peak-calling Pipeline:

```
#Align fastq files to BL6/CAST indexed assembly.

bwa index -a bwtsw /genome/mm10.fa
bwa mem /genome/mm10.fa CTCF_ChIPseq_library.p1.fq
CTCF_ChIPseq_library.p2.fq > CTCF_ChIPseq_library.sam

# Fetch uniquely mapped reads, outputting filtered SAM alignment files

fgrep -w "XT:A:U" CTCF_ChIPseq_library.sam >
CTCF_ChIPseq_library_filtered.sam

#Convert SAM file to BAM, followed by sorting the BAM file

samtools view -bSo CTCF_ChIPseq_library.bam
CTCF_ChIPseq_library_filtered.sam
samtools sort CTCF_ChIPseq_library.bam CTCF_ChIPseq_library.sorted

#Index sorted BAM file

samtools index CTCF_ChIPseq_library.sorted.bam

#NOTE: repeat all steps above for control/input libraries

#Peak-calling using MACS2 with default parameters

macs2 callpeak -t CTCF_ChIPseq_library.sorted.bam -c
CTCF_ChIPseq_input.sorted.bam -f 'BAM' -g 'mm' -n CTCF_ChIPseq_library
-B --call-summits -p 0.001
```

Quality Control, Alignment and Peak-calling Pipeline (cis/trans):

```
#Trim Library using min Phred score of 33, performing a sliding window
trimming of window size 20, average quality of 30 and removing reads
that fall below 40 minimal length

java -jar trimmomatic-0.30.jar SE -phred33 CTCF_ChIPseq_library.fq
CTCF_ChIPseq_library.fq_trim SLIDINGWINDOW:20:30 MINLEN:40

#Align trimmed fastq files to joint assembly. First find SA coords of
trim_fa (6 threads, 2 maximum mismatches) and output to .sai. Next,
align to assembly using .sai and trim_fa, output to trim.m2.samtemp.

bwa aln -t 6 -n 2 /genome/mm10 CTCF_ChIPseq_library.fq_trim >
CTCF_ChIPseq_library.fq_trim.sai
bwa samse CTCF_ChIPseq_library.fq_trim.sai
CTCF_ChIPseq_library.fq_trim  >
CTCF_ChIPseq_library.fq_trim.m2.samtemp

# Fetch uniquely mapped reads, outputting SAM alignment files

fgrep -w "XT:A:U" CTCF_ChIPseq_library.fq_trim.m2.samtemp >
CTCF_ChIPseq_library.fq_trim.m2.sam

#Convert SAM file to BAM, followed by sorting the BAM file

samtools view -S -b CTCF_ChIPseq_library.fq_trim.m2.sam >
CTCF_ChIPseq_library.fq_trim.m2.bam
samtools sort CTCF_ChIPseq_library.fq_trim.m2.bam
CTCF_ChIPseq_library.fq_trim.m2

#mpileup used to count the number of reads that overlapped each base
of the ref genome. Maximum depth of 100,000, No probabilistic
realignment for the computation of base alignment quality (BAQ) or
Minimum base quality.

samtools mpileup -I  -d 100000 -BQ0 -f /genome/mm10.fa
CTCF_ChIPseq_library.fq_trim.m2.bam >
CTCF_ChIPseq_library.fq_trim.m2.mpileup

#Index sorted BAM file

samtools index CTCF_ChIPseq_library.fq_trim.m2.bam

#Peak-calling using MACS2 with default parameters
macs2 callpeak -t CTCF_ChIPseq_library.fq_trim.m2.bam -c
do3072_input_liver_none_mm9BL6xCAST82602.0_CRI01.fq.gz_trim.m2.bam -f
'BAM' -g 'mm' -n CTCF_ChIPseq_library.fq_trim.m2.sam_macs
```

#This R code runs statistical model fitting for category-assignment of cis/trans regulatory variants in TF binding sites. The code below uses an example of CTCF in 2 biological replicates. The R code sources the two helper functions f1_functions.R and cistrans.R.

f1_functions.R

```
library(DESeq)

#This function normalizes F0 libraries based on library size using
DESeq, producing a fitted dispersion parameter (r) for each SNV/site

f0norm = function(pdata, pcondition){
     pcds = newCountDataSet(pdata, pcondition)
     pcds = estimateSizeFactors(pcds)
     pcds = estimateDispersions(pcds, fitType='local')
     r = 1/fitInfo(pcds)$fittedDispEsts
     normcounts = cbind(as.matrix(counts(pcds, normalized=TRUE)), r)
     return(normcounts)
}

#This function normalizes F1 libraries based on library size using
DESeq.

f1norm = function(castf, b6f){
     summed = castf + b6f
     b6f = subset(b6f, row.names(b6f) %in% row.names(summed))
     castf = subset(castf, row.names(castf) %in% row.names(summed))
     fract = b6f / (castf + b6f)
     condition = factor(rep('a', ncol(summed)))
     cds = newCountDataSet(summed, condition)
     cds = estimateSizeFactors(cds)
     normcounts = as.matrix(counts(cds, normalized=TRUE))
     normb6 = normcounts * fract
     normcast = normcounts * (1 - fract)
     normb6[is.na(normb6)] = 0
     normcast[is.na(normcast)] = 0
     return(list(normb6, normcast))
}


#This function estimates the Bayesian Information Criteria (BIC)

BIC = function(loglik , npar , nobs){
     bic = -2*loglik + npar*log(nobs)
}
```

cistrans.R

```
#Cis/trans function to model counts distribution between F0 and F1. F1
are modelled on a beta-binomial distribution, and F0 counts are
modelled on negative binomial distributions.

f0f1_func = function(x, y, n, z, p1, p2, a, b, r) {
     sum(lchoose(n,z)+lbeta(z+a,n-z+b)-lbeta(a,b)) +
     sum(lgamma(r + x) - lfactorial(x) - lgamma(r) + x*log(p1) +
r*log(1-p1)) +
     sum(lgamma(r + y) - lfactorial(y) - lgamma(r) + y*log(p2) +
r*log(1-p2))
}

#Parameter estimation of conserved scenario with 2 free parameters.

cons_func = function(params, x, y, n, z, r) {
     p1 = params[1]
     p2 = p1
     a = params[2]
     b = a
     f0f1_func(x, y, n, z, p1, p2, a, b, r)
}

#Parameter estimation of cis scenario with 3 free parameters.

cis_func = function(params, x, y, n, z, r) {
     p1 = params[1]
     p2 = params[2]
     a = params[3]
     b = (a * ((p1/(1-p1)) + (p2/(1-p2))) / (p1/(1-p1))) -a
     f0f1_func(x, y, n, z, p1, p2, a, b, r)
}

#Parameter estimation of trans scenario with 3 free parameters.

trans_func = function(params, x, y, n, z, r) {
     p1 = params[1]
     p2 = params[2]
     a = params[3]
     b = a
     f0f1_func(x, y, n, z, p1, p2, a, b, r)
}

#Parameter estimation of cistrans scenario with 4 free parameters.

cistrans_func = function(params, x, y, n, z, r) {
     p1 = params[1]
     p2 = params[2]
     a = params[3]
     b = params[4]
     f0f1_func(x, y, n, z, p1, p2, a, b, r)
}

#Statistical modelling of each scenario based on F0 & F1 normalised
counts, dispersion and free parameters.
```

```
cistrans_ml = function(x, y, n, z, p1, p2, a, b, r) {
       res = c()
       for(i in 1:nrow(x)) {
       res = c(res, f0f1_func(x[i,], y[i,], n[i,], z[i,], p1[i],
p2[i], a[i], b[i], r[i]))
     }
     return(res)
}
```

---

cistrans_cat_assignment.R

```
library(DESeq)

source("f1_functions.R")
source("cistrans.R")

# Load reads pileups for SNVs/sites for all F0 & F1 libraries
filtered = read.delim("CTCF_reads_df")

# Normalise reads from F0 libraries
df=filtered[,c(2:5)]
row.names(df)=filtered[,1]
condition = factor(c('p1','p1' ,'p2','p2'))
df = f0norm(df, condition)

# Normalise reads from F1 libraries
df1_cast=filtered[,c(6:9)]
row.names(df1_cast)=filtered$chr
df1_b6=filtered[,c(10:13)]
row.names(df1_b6)=filtered$chr
x = f1norm(df1_cast, df1_b6)
norm_b6 = as.data.frame(x[1])
norm_cast = as.data.frame(x[2])
colnames(norm_cast) = colnames(df1_cast)
f1 = cbind(norm_b6 , norm_cast)

# Combine normalised counts for F0 & F1 libraries
df = merge(df, f1, by='row.names')
row.names(df) = df$Row.names
df$Row.names = NULL

# Set normalised counts for cis/trans category-assignment
#x=F0 BL6, y=F0 CAST, n= F1 B6L+ F1 CAST, z=F1 BL6

x=as.matrix(df[,c(1:2)])
y=as.matrix(df[,c(3:4)])
n=as.matrix(df[,c(6:9)]) + as.matrix(df[,c(10:13)])
z=as.matrix(df[,c(6:9)])

x=round(x)
y=round(y)
z=round(z)
n=round(n)
```

```
r = df$r

#For each SNV/site, x, y are the parental data, and n, z are the
hybrid offspring data

#Parameter estimation for the (conserved) scenario. initcons = initial
values for parameter optimisation.


res = c()
initcons = c(0.5, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcons, cons_func, gr=NULL, x[i,], y[i,],n[i,],
z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8, 1e-8), upper=c(0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "alpha")
p1 = res[,"p1"]
alpha = res[,"alpha"]

cons = cistrans_ml(x, y, n, z, p1, p1, alpha, alpha, r)
df = cbind(df, cons)

#Parameter estimation for the (cis) scenario. initcons = initial
values for parameter optimisation.

res = c()
initcis = c(0.5, 0.5, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcis, cis_func, gr=NULL, x[i,], y[i,], n[i,],
z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8),
upper=c(0.9999,0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "p2", "alpha")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]
beta = (alpha*p2*(1-p1))/(p1*(1-p2))

cis = cistrans_ml(x, y, n, z, p1, p2, alpha, beta, r)
df = cbind(df, cis)

#Parameter estimation for the (trans) scenario. initcons = initial
values for parameter optimisation.

res = c()
inittrans =c(0.5, 0.5, 1)
for(i in 1:nrow(x)) {
```

```
     ores = optim(inittrans, trans_func, gr=NULL, x[i,], y[i,], n[i,],
z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8),
upper=c(0.9999,0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "p2", "alpha")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]

trans = cistrans_ml(x, y, n, z, p1, p2, alpha, alpha, r)
df = cbind(df, trans)

#Parameter estimation for the (cistrans) scenario. initcons = initial
values for parameter optimisation.

res = c()
initcistrans =c(0.5, 0.5, 1, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcistrans, cistrans_func, gr=NULL, x[i,], y[i,],
n[i,], z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8,1e-8),
     upper=c(0.9999,0.9999,1e6,1e6), control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "p2", "alpha", "beta")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]
beta = res[,"beta"]

cistrans = cistrans_ml(x, y, n, z, p1, p2, alpha, beta, r)
df = cbind(df, cistrans)

#Estimation of Bayesian Information Criteria (BIC). npar = Number of
parameters. nobs = Number of samples/libraries.

BIC = function(loglik , npar , nobs){
     bic = -2*loglik + npar*log(nobs)
}

#BIC estimation for each scenario. BIC tests all possible models:
#cons has 2 free parameters
#cis and trans both have 3 free parameters
#cistrans has 4 free parameters

df$cons_bic = BIC(df$cons, 2 , 2)
df$cis_bic = BIC(df$cis, 3 , 2)
df$trans_bic = BIC(df$trans, 3 , 2)
df$cistrans_bic = BIC(df$cistrans, 4 ,2)
```

```
#Assigning BIC values to all four models tested above

inds = apply(df[,c("cons_bic" , "cis_bic" , "trans_bic" ,
"cistrans_bic")] , 1 , function(x) which(x==min(x), arr.ind=TRUE))
cat = c("cons" , "cis" , "trans" , "cistrans")[inds]
df = cbind(df, cat)
df$minBIC = apply(df[,c("cons_bic" , "cis_bic" , "trans_bic" ,
"cistrans_bic")] , 1 , min)
minn = function(n) function(x) order(x, decreasing = FALSE)[n]
df$secondlowestBIC = apply(df[,c("cons_bic" , "cis_bic" , "trans_bic"
, "cistrans_bic")] , 1 , function(x) x[minn(2)(x)])

df$dif_bic =df$secondlowestBIC - df$minBIC

write.table(df, "rdata/CTCF_df_x2_cistrans", sep="\t", row.names=TRUE,
quote=FALSE, col.names=TRUE)
```

#This R code randomly subsamples n (here n=2) number of replicates out of 6, then runs statistical model fitting for category-assignment of cis/trans regulatory variants in TF binding sites. The code below uses an example of 2 biological replicates from a TF of interest. The shell script generate parallel jobs for the random subsampling. The R code sources the two helper functions f1_functions.R and cistrans.R detailed above.

generate_jobs.sh

```
#!/bin/bash
if [[ "$#" -ne 3 ]]; then
echo "Error: Usage ./generate_jobs.sh <jobs> <runs> <input file>"
exit 1
fi

JOBS="$1"
RUNS="$2"
INPUT_FILE="$3"

for ((i=1; i<=$JOBS; i++))
do
echo "Running job: $i ($RUNS runs using $INPUT_FILE for input)"
bsub rscript Random_Subsampling.r $i $RUNS $INPUT_FILE
done
```

Random_Subsamplingx2.r

```
setwd("/Subsampling/x2/")

library(DESeq)

source("f1_functions.R")
source("cistrans.R")

# Get a list of args from the command line.
args = commandArgs(trailingOnly=TRUE)

# To run the script, three arguments are required:
#
# job: the number of the job being run
# runs: the number of runs to go through
# input file: the name of the input file
#
# Example: rscript Random_Subsampling.r 5 200 CEBPA_reads_df
#
if (length(args) != 3) {
      stop("Usage: rscript Random_Subsampling.r <job> <runs> <input
file>")
}
```

```
# Load the variables from the arguments.
job_number = args[1]
run_count = args[2]
input_file_name = args[3]

# Generate the output file name from the arguments.
# The output file name has the format:
# <input file name>_<job>_<runs>
#
# Example: CEBPA_reads_df_5_200
#
output_file_name = paste(input_file_name, job_number, run_count,
sep="_")

subsampling_stats = data.frame(cis= integer(0), cistrans = integer(0),
cons= integer(0), trans= integer(0))

# Load reads pileups for SNVs/sites for all F0 & F1 libraries
filtered = read.delim(input_file_name)


for(i in 1:run_count) {
print(i)

start_time = Sys.time()

#Randomly subsample n from 6 replicate (in this example n=2)
xb6 = sample(2:7, 2)
xcast = sample(8:13, 2)

#Randomly subsample from 6 F1 replicate, selecting both alleles from
the same hybrid offspring.
xf1_cast_i = sample(14:19,2)
xf1_cast_r = sample(20:25,2)
xf1_b6_i = xf1_cast_i + 12
xf1_b6_r = xf1_cast_r + 12

tryCatch(
  {
  # Normalise reads from F0 libraries
  df=filtered[,c(xb6,xcast)]
  row.names(df)=filtered[,1]
  condition = factor(c('p1','p1','p2','p2'))
  df = f0norm(df, condition)

  # Normalise reads from F1 libraries
  df1_cast=filtered[,c(xf1_cast_i, xf1_cast_r)]
  row.names(df1_cast)=filtered$chr
  df1_b6=filtered[,c(xf1_b6_i, xf1_b6_r)]
  row.names(df1_b6)=filtered$chr
  x = f1norm(df1_cast, df1_b6)
  norm_b6 = as.data.frame(x[1])
  norm_cast = as.data.frame(x[2])
  colnames(norm_cast) = colnames(df1_cast)
  f1 = cbind(norm_b6 , norm_cast)
```

```
# Combine normalised counts for F0 & F1 libraries
df = merge(df, f1, by='row.names')
row.names(df) = df$Row.names
df$Row.names = NULL

# Set normalised counts for cis/trans category-assignment
#x=F0 BL6, y=F0 CAST, n= F1 B6L+ F1 CAST, z=F1 BL6

x=as.matrix(df[,c(1:2)])
y=as.matrix(df[,c(3:4)])
n=as.matrix(df[,c(6:9)]) + as.matrix(df[,c(10:13)])
z=as.matrix(df[,c(6:9)])

x=round(x)
y=round(y)
z=round(z)
n=round(n)
r = df$r

#For each SNV/site, x, y are the parental data, and n, z are the
hybrid offspring data

#Parameter estimation for the (conserved) scenario. initcons =
initial values for parameter optimisation.

res = c()
initcons = c(0.5, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcons, cons_func, gr=NULL, x[i,], y[i,],n[i,],
z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8, 1e-8),
upper=c(0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "alpha")
p1 = res[,"p1"]
alpha = res[,"alpha"]

cons = cistrans_ml(x, y, n, z, p1, p1, alpha, alpha, r)
df = cbind(df, cons)

#Parameter estimation for the (cis) scenario. initcons = initial
values for parameter optimisation.

res = c()
initcis = c(0.5, 0.5, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcis, cis_func, gr=NULL, x[i,], y[i,], n[i,],
z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8),
upper=c(0.9999,0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}
```

```
colnames(res) = c("p1", "p2", "alpha")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]
beta = (alpha*p2*(1-p1))/(p1*(1-p2))

cis = cistrans_ml(x, y, n, z, p1, p2, alpha, beta, r)
df = cbind(df, cis)

#Parameter estimation for the (conserved) scenario. initcons =
initial values for parameter optimisation.

res = c()
inittrans =c(0.5, 0.5, 1)
for(i in 1:nrow(x)) {
     ores = optim(inittrans, trans_func, gr=NULL, x[i,], y[i,],
n[i,], z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8),
upper=c(0.9999,0.9999,1e6),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "p2", "alpha")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]

trans = cistrans_ml(x, y, n, z, p1, p2, alpha, alpha, r)
df = cbind(df, trans)

#Parameter estimation for the (conserved) scenario. initcons =
initial values for parameter optimisation.

res = c()
initcistrans =c(0.5, 0.5, 1, 1)
for(i in 1:nrow(x)) {
     ores = optim(initcistrans, cistrans_func, gr=NULL, x[i,],
y[i,], n[i,], z[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8,1e-8,1e-8),
     upper=c(0.9999,0.9999,1e6,1e6), control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("p1", "p2", "alpha", "beta")
p1 = res[,"p1"]
p2 = res[,"p2"]
alpha = res[,"alpha"]
beta = res[,"beta"]

cistrans = cistrans_ml(x, y, n, z, p1, p2, alpha, beta, r)
df = cbind(df, cistrans)

#Estimation of Bayesian Information Criteria (BIC). npar = Number of
parameters. nobs = Number of samples/libraries.
```

```
   BIC = function(loglik , npar , nobs){

   bic = -2*loglik + npar*log(nobs)
   }

   #BIC estimation for each scenario. BIC tests all possible models:
   #cons has 2 free parameters
   #cis and trans both have 3 free parameters
   #cistrans has 4 free parameters

   df$cons_bic = BIC(df$cons, 2 , 2)
   df$cis_bic = BIC(df$cis, 3 , 2)
   df$trans_bic = BIC(df$trans, 3 , 2)
   df$cistrans_bic = BIC(df$cistrans, 4 , 2)

   #Assigning BIC values to all four models tested above
   inds = apply(df[,c("cons_bic" , "cis_bic" , "trans_bic" ,
   "cistrans_bic")] , 1 , function(x) which(x==min(x), arr.ind=TRUE))
   cat = c("cons" , "cis" , "trans" , "cistrans")[inds]
   df = cbind(df, cat)
   df$minBIC = apply(df[,c("cons_bic" , "cis_bic" , "trans_bic" ,
   "cistrans_bic")] , 1 , min)
   minn = function(n) function(x) order(x, decreasing = FALSE)[n]
   df$secondlowestBIC = apply(df[,c("cons_bic" , "cis_bic" ,
   "trans_bic" , "cistrans_bic")] , 1 , function(x) x[minn(2)(x)])

   df$dif_bic =df$secondlowestBIC - df$minBIC

   summary_stat = as.data.frame(as.list(summary(df$cat)))

   subsampling_stats = rbind(subsampling_stats,summary_stat)
   },
error = function(error_condition) {
     print(error_condition)
     print("------")
     print(xb6)
     print(xcast)
     print(xf1_cast_i)
     print(xf1_cast_r)
     print(xf1_b6_i)
     print(xf1_b6_r)
     print("------")
},
finally={
     end_time = Sys.time()
     print(end_time - start_time)
}
)

}

write.table(subsampling_stats,    output_file_name,    row.names=TRUE,
sep="\t", quote=FALSE, col.names=TRUE)
```

#This R code runs a statistical model fitting for determining TF binding sites modes of inheritance. The code below uses an example of CTCF in 2 biological replicates.

inheritance_functions.R

```
#Mode of inheritance assignment function modelling counts distribution
between F0 and F1. All counts are modelled on negative binomial
distributions.
inheritance_func = function(xmax, xmin, yinh, inh_p1, inh_p2, inh_p3,
r) {
     sum(lgamma(r + xmax) - lfactorial(xmax) - lgamma(r)
+xmax*log(inh_p1) + r*log(1-inh_p1)) +
     sum(lgamma(r + xmin) - lfactorial(xmin) - lgamma(r)
+xmin*log(inh_p2) + r*log(1-inh_p2)) +
     sum(lgamma(r + yinh) - lfactorial(yinh) - lgamma(r)
+yinh*log(inh_p3) + r*log(1-inh_p3))
}

#Parameter estimation of additive mode with 3 free parameters.
additive_func = function(params, xmax, xmin, yinh, r) {
     inh_p1 = params[1]
     inh_p2 = params[2]
     inh_p3 = params[3]
     inheritance_func(xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
}

#Parameter estimation of dominant mode with 2 free parameters (3rd
parameter = parameter 2).
dom1_func = function(params, xmax, xmin, yinh, r) {
     inh_p1 = params[1]
     inh_p2 = params[2]
     inh_p3 = inh_p2
     inheritance_func(xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
}

#Parameter estimation of dominant mode with 2 free parameters (3rd
parameter = parameter 1).
dom2_func = function(params, xmax, xmin, yinh, r) {
     inh_p1 = params[1]
     inh_p2 = params[2]
     inh_p3 = inh_p1
     inheritance_func (xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
}

#Parameter estimation of the excluded set with 1 free parameters (3rd
parameter = parameter 1 & 2 simultaneously).
exclude_func = function(params, xmax, xmin, yinh, r) {
     inh_p1 = params[1]
     inh_p2 = inh_p1
     inh_p3 = inh_p2
     inheritance_func (xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
```

```
}

#Statistical modelling of each mode based on F0 & F1 counts,
dispersion and free parameters.
inheritance_ml = function(xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
{
     res = c()
     for(i in 1:nrow(xmax)) {
     res = c(res, inheritance_func(xmax[i,], xmin[i,], yinh[i,],
     inh_p1[i], inh_p2[i], inh_p3[i], r[i]))
}
return(res)
}
```

---

inheritance_stat_modelling.R

```
source("inheritance_functions.R")

# Load category-assigned, normalised counts for SNVs/sites in all F0 &
F1 libraries
CTCF_analysis_x2_df = read.table("rdata/CTCF_df_x2_cistrans.txt", h=T)

#Filter by differential BIC value >=1, sub-setting cis and trans
categories for further analysis
CTCF_x2_inh_df = subset(CTCF_analysis_x2_df, dif_bic >= 1)
CTCF_x2_inh_df = subset(CTCF_x2_inh_df, cat == "cis" | cat == "trans")

#Calculating the means of F0 between parental strains
CTCF_x2_inh_df$f0_b6_av = apply(CTCF_x2_inh_df[,3:4], 1, mean)
CTCF_x2_inh_df$f0_cast_av = apply(CTCF_x2_inh_df[,5:6], 1, mean)
CTCF_x2_inh_df$diff_f0 = abs(CTCF_x2_inh_df$f0_b6_av-
CTCF_x2_inh_df$f0_cast_av)

#Calculating the median binding intensity in both strains of F0
CTCF_x2_inh_df$f0_b6_med = apply(CTCF_x2_inh_df[,3:4], 1, median)
CTCF_x2_inh_df$f0_cast_med = apply(CTCF_x2_inh_df[,5:6], 1, median)

#Calculating the median binding intensity in both alleles of F1, and
summing them.
CTCF_x2_inh_df$f1_total_av = (apply(CTCF_x2_inh_df[,8:11], 1,
median))+(apply(CTCF_x2_inh_df[,12:15], 1, median))

#Determining the F0 parent with the higher/lower median binding
intensity
CTCF_x2_inh_df$xmax =
apply(CTCF_x2_inh_df[,c("f0_b6_med","f0_cast_med")], 1, max)
CTCF_x2_inh_df$xmin =
apply(CTCF_x2_inh_df[,c("f0_b6_med","f0_cast_med")], 1, min)

#For the F0 paerent with the higher median binding intensity, retrieve
the F0 normalised counts for that parent to use in model fitting.

x_max = data.frame()
for(i in 1:nrow(CTCF_x2_inh_df)) {
```

```
     if(CTCF_x2_inh_df[i,31] > CTCF_x2_inh_df[i,32]){
          xmax = CTCF_x2_inh_df[i,3:4]
          names(xmax) = c("x1","x2)
     } else{
          xmax = CTCF_x2_inh_df[i,5:6]
          names(xmax) = c("x1","x2")
}
x_max = rbind(x_max, xmax)
}

#For the F0 parent with the higher median binding intensity, retrieve
the F0 normalised counts for that parent to use in model fitting.

x_min = data.frame()
for(i in 1:nrow(CTCF_x2_inh_df)) {
     if(CTCF_x2_inh_df[i,31] < CTCF_x2_inh_df[i,32]){
          xmin = CTCF_x2_inh_df[i,3:4]
          names(xmin) = c("x1","x2")
     } else{
          xmin = CTCF_x2_inh_df[i,5:6]
          names(xmin) = c("x1","x2")
     }
     x_min = rbind(x_min, xmin)
}

#Set variables for assigning modes of inheritance.
#xmax=F0 with highest signal, xmin=F0 with lowest signal, yinh= total
allelic signal in F1

xmax=as.matrix(x_max)
xmin=as.matrix(x_min)
yinh=(as.matrix(CTCF_x2_inh_df[,c(8:11)]) +
as.matrix(CTCF_x2_inh_df[,c(12:15)]))/2

xmax=round(xmax)
xmin=round(xmin)
yinh=round(yinh)

r = CTCF_x2_inh_df$r

#For each SNV/site, xmax and xmin are the parental data, and yinh are
the hybrid offspring data

#Parameter estimation for the additive mode. initcons = initial values
for parameter optimisation.

res = c()
init_add = c(0.5, 0.5, 0.5)
for(i in 1:nrow(xmax)) {
     ores = optim(init_add, additive_func, gr=NULL, xmax[i,],
xmin[i,],yinh[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8, 1e-8, 1e-8),
upper=c(0.9999,0.9999,0.9999),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}
```

```
colnames(res) = c("inh_p1", "inh_p2", "inh_p3")
inh_p1 = res[,"inh_p1"]
inh_p2 = res[,"inh_p2"]
inh_p3 = res[,"inh_p3"]

additive = inheritance_ml(xmax, xmin, yinh, inh_p1, inh_p2, inh_p3, r)
CTCF_x2_inh_df = cbind(CTCF_x2_inh_df, additive)

# Parameter estimation of dominant mode with 2 free parameters (3rd
parameter = parameter 2).

res = c()
init_dom1 = c(0.5, 0.5)
for(i in 1:nrow(xmax)) {
     ores = optim(init_dom1, dom1_func, gr=NULL, xmax[i,], xmin[i,],
yinh[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8), upper=c(0.9999,0.9999),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("inh_p1", "inh_p2")
inh_p1 = res[,"inh_p1"]
inh_p2 = res[,"inh_p2"]

dom1 = inheritance_ml(xmax, xmin, yinh, inh_p1, inh_p2, inh_p2, r)
CTCF_x2_inh_df = cbind(CTCF_x2_inh_df, dom1)


# Parameter estimation of dominant mode with 2 free parameters (3rd
parameter = parameter 1).

res = c()
init_hidom =c(0.5, 0.5)
for(i in 1:nrow(xmax)) {
     ores = optim(init_hidom, dom2_func, gr=NULL, xmax[i,], xmin[i,],
yinh[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8), upper=c(0.9999,0.9999),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("inh_p1", "inh_p2")
inh_p1 = res[,"inh_p1"]
inh_p2 = res[,"inh_p2"]

dom2 = inheritance_ml(xmax, xmin, yinh, inh_p1, inh_p2, inh_p1, r)
CTCF_x2_inh_df = cbind(CTCF_x2_inh_df, dom2)

#Parameter estimation of the excluded set with 1 free parameters (3rd
parameter = parameter 1 & 2 simultaneously).

res = c()
init_exc =c(0.5)
for(i in 1:nrow(xmax)) {
```

```
     ores = optim(init_exc, exclude_func, gr=NULL, xmax[i,], xmin[i,],
yinh[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8), upper=c(0.9999), control =
list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("inh_p1")
inh_p1 = res[,"inh_p1"]

exclude = inheritance_ml(xmax, xmin, yinh, inh_p1, inh_p1, inh_p1, r)
CTCF_x2_inh_df = cbind(CTCF_x2_inh_df, exclude)

#Estimation of Bayesian Information Criteria (BIC). npar = Number of
parameters. nobs = Number of samples/libraries.

BIC = function(loglik , npar , nobs){

     bic = -2*loglik + npar*log(nobs)
}

#BIC estimation for each inheritance mode. BIC tests all possible
models:
#additive has 3 free parameters
#both dominant modes have 2 free parameters
#excluded sites have 1 free parameters

CTCF_x2_inh_df$addit_bic = BIC(CTCF_x2_inh_df$additive, 3 ,2)
CTCF_x2_inh_df$dom1_bic = BIC(CTCF_x2_inh_df$dom1, 2 ,2)
CTCF_x2_inh_df$hidom_bic = BIC(CTCF_x2_inh_df$dom2, 2 ,2)
CTCF_x2_inh_df$exclu_bic = BIC(CTCF_x2_inh_df$exclude, 1 ,2)

#Assigning BIC values to all modes tested above
inds = apply(CTCF_x2_inh_df[,c("addit_bic" , "dom1_bic" , "hidom_bic"
, "exclu_bic")] , 1 , function(x) which(x==min(x), arr.ind=TRUE))
inh_cat = c("additive" , "dom1" , "dom2" , "exclude")[inds]
CTCF_x2_inh_df = cbind(CTCF_x2_inh_df, inh_cat)
CTCF_x2_inh_df$inh_minBIC = apply(CTCF_x2_inh_df[,c("addit_bic" ,
"dom1_bic" , "hidom_bic" , "exclu_bic")] , 1 , min)

minn = function(n) function(x) order(x, decreasing = FALSE)[n]
CTCF_x2_inh_df$inh_secondlowestBIC =
apply(CTCF_x2_inh_df[,c("addit_bic" , "dom1_bic" , "hidom_bic" ,
"exclu_bic")] , 1 , function(x) x[minn(2)(x)])
CTCF_x2_inh_df$inh_dif_bic =CTCF_x2_inh_df$inh_secondlowestBIC -
CTCF_x2_inh_df$inh_minBIC

summary(CTCF_x2_inh_df$inh_cat[CTCF_x2_inh_df$cat == "cis"])
summary(CTCF_x2_inh_df$inh_cat[CTCF_x2_inh_df$cat == "trans"])

write.table(CTCF_x2_inh_df, "rdata/CTCF_x2_inh_df", sep="\t", row.names
= TRUE, quote=FALSE, col.names=TRUE)
```

#This R code runs a statistical model fitting for lineage-specific TF binding sites. The code below uses an example of CTCF in 2 biological replicates.

Lineage_spec_functions.R

```
#Function modelling counts distribution in lineage-specific binding
between F0 and F1. All counts are modelled on negative binomial
distributions.

lineage_func = function(xi, yi, lin_p1, lin_p2, r) {
     sum(lgamma(r + xi) - lfactorial(xi) - lgamma(r) +xi*log(lin_p1) +
     r*log(1-lin_p1)) +
     sum(lgamma(r + yi) - lfactorial(yi) - lgamma(r) +yi*log(lin_p2) +
     r*log(1-lin_p2))
}

#Parameter estimation of lineage-specific cis scenario with 1 free
parameter.

lin_cis_func = function(params, xi, yi, r) {
     lin_p1 = params[1]
     lin_p2 = lin_p1
     lineage_func(xi, yi, lin_p1, lin_p2, r)
}

#Parameter estimation of lineage-specific cistrans scenario with 2
free parameters.

lin_cistrans_func = function(params, xi, yi, r) {
     lin_p1 = params[1]
     lin_p2 = params[2]
     lineage_func(xi, yi, lin_p1, lin_p2, r)
}

#Statistical modelling of each scenario based on F0 & F1 counts,
dispersion and free parameters.

lineage_ml = function(xi, yi, lin_p1, lin_p2, r) {
     res = c()
     for(i in 1:nrow(xi)) {
          res = c(res, lineage_func(xi[i,], yi[i,], lin_p1[i],
          lin_p2[i], r[i]))
     }
     return(res)
}
```

Lineage_spec_stat_modelling.R

```
source("../lineage_functions.R")
```

```
# Load category-assigned, normalised counts for SNVs/sites in all F0 &
F1 libraries
CTCF_analysis_x2_df = read.table("rdata/CTCF_df_x2_cistrans.txt", h=T)

#Calculating the ratios of F0 and F1
CTCF_analysis_x2_df$f0_ratio = apply(CTCF_analysis_x2_df[,3:4], 1,
sum)/ apply(CTCF_analysis_x2_df[,3:6], 1, sum)
CTCF_analysis_x2_df$f1_ratio = apply(CTCF_analysis_x2_df[,8:11], 1,
sum)/ apply(CTCF_analysis_x2_df[,8:15], 1, sum)

#Calculating the means of F0 between parental strains
CTCF_analysis_x2_df $f0_b6_av = apply(CTCF_analysis_x2_df[,3:4], 1,
mean)
CTCF_analysis_x2_df $f0_cast_av = apply(CTCF_analysis_x2_df[,5:6], 1,
mean)

#Calculating the mean binding intensity in both alleles of F1, and
summing them.
CTCF_analysis_x2_df$f1_total_av = (apply(CTCF_analysis_x2_df[,8:11],
1, median))+(apply(CTCF_analysis_x2_df[,12:15], 1, mean))

#Determine lineage-specificity based on F0 and F1 ratios between
parents and offspring for further analysis

CTCF_x2_linspec_df = filter(CTCF_analysis_x2_df, f0_ratio > 0.95 &
f1_ratio > 0.95 | f0_ratio < 0.05 & f1_ratio < 0.05)

#Determining the F0 parent of lineage-specific binding.

CTCF_x2_linspec_df$xi =
apply(CTCF_x2_linspec_df[,c("f0_b6_av","f0_cast_av")], 1, max)


#For the F0 parent of lineage-specific binding, retrieve the F0
normalised counts for that parent to use in model fitting.

x_i = data.frame()
for(i in 1:nrow(CTCF_x2_linspec_df)) {
     if(CTCF_x2_linspec_df[i,30] > CTCF_x2_linspec_df[i,31]){
          xi = CTCF_x2_linspec_df[i,3:4]
          names(xi) = c("x1","x2)
     } else{
          xi = CTCF_x2_linspec_df[i,5:6]
          names(xi) = c("x1","x2")
}
x_i = rbind(x_i, xi)
}

#Set variables for investigating lineage-specificity of binding.
#xi=F0 with lineage-spcific binding, yi= total allelic signal in F1

xi=as.matrix(x_i)
yi=(as.matrix(CTCF_x2_linspec_df[,c(8:11)]) +
as.matrix(CTCF_x2_linspec_df[,c(12:15)]))
```

```
xi=round(xi)
yi=round(2*yi)
r = CTCF_x2_linspec_df$r

#For each SNV/site, xi are the parental data, and yi are the hybrid
offspring data

#Parameter estimation for the cis scenario. initcons = initial values
for parameter optimisation.

res = c()
init_lin_cis =c(0.5)
for(i in 1:nrow(xi)) {
     ores = optim(init_lin_cis, lin_cis_func, gr=NULL, xi[i,], yi[i,],
     r[i],
     method = "L-BFGS-B", lower = c(1e-8), upper=c(0.9999), control =
     list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("lin_p1")
lin_p1 = res[,"lin_p1"]

lin_cis = lineage_ml(xi, yi, lin_p1, lin_p1, r)
CTCF_x2_linspec_df = cbind(CTCF_x2_linspec_df, lin_cis)

#Parameter estimation for the cistrans scenario.

res = c()
init_lin_cistrans = c(0.5, 0.5)
for(i in 1:nrow(xi)) {
     ores = optim(init_lin_cistrans, lin_cistrans_func, gr=NULL,
     xi[i,], yi[i,], r[i],
     method = "L-BFGS-B", lower = c(1e-8,1e-8), upper=c(0.9999,0.9999),
     control = list(fnscale=-1))
     res = rbind(res, ores$par)
}

colnames(res) = c("lin_p1", "lin_p2")
lin_p1 = res[,"lin_p1"]
lin_p2 = res[,"lin_p2"]

lin_cistrans = lineage_ml(xi, yi, lin_p1, lin_p2, r)
CTCF_x2_linspec_df = cbind(CTCF_x2_linspec_df, lin_cistrans)

#Estimation of Bayesian Information Criteria (BIC). npar = Number of
parameters. nobs = Number of samples/libraries.

BIC = function(loglik , npar , nobs){

     bic = -2*loglik + npar*log(nobs)
}

#BIC estimation for each scenario. BIC tests two possible models:
#cis have 1 free parameter
```

```
#cistrans have 2 free parameters

CTCF_x2_linspec_df$lin_cis_bic = BIC(CTCF_x2_linspec_df$lin_cis, 1 ,2)
CTCF_x2_linspec_df$lin_cistrans_bic =
BIC(CTCF_x2_linspec_df$lin_cistrans, 2 ,2)


#Assigning BIC values to all scenarios tested above

inds = apply(CTCF_x2_linspec_df[,c("lin_cis_bic" ,
"lin_cistrans_bic")] , 1 , function(x) which(x==min(x), arr.ind=TRUE))
lin_cat = c("lin_cis" , "lin_cistrans")[inds]
CTCF_x2_linspec_df = cbind(CTCF_x2_linspec_df, lin_cat)
CTCF_x2_linspec_df$lin_minBIC =
apply(CTCF_x2_linspec_df[,c("lin_cis_bic" , "lin_cistrans_bic")] , 1 ,
min)

minn = function(n) function(x) order(x, decreasing = FALSE)[n]
CTCF_x2_linspec_df$lin_secondlowestBIC =
apply(CTCF_x2_linspec_df[,c("lin_cis_bic" , "lin_cistrans_bic")] , 1 ,
function(x) x[minn(2)(x)])

CTCF_x2_linspec_df$lin_dif_bic =CTCF_x2_linspec_df$lin_secondlowestBIC
- CTCF_x2_linspec_df$lin_minBIC

write.table(CTCF_x2_linspec_df, "rdata/ CTCF_x2_linspec_df", sep="\t",
row.names = TRUE, quote=FALSE, col.names=TRUE)
```

#This Python code with its helper function (helpers.py) takes as in input a csv filer with four columns. The first two columns describe the chromosome number and position of each cis SNV. The third and fourth columns describe the chromosome number and position of all SNVs. The function anchors each cis SNV and searches for other SNVs in downstream incremental genomic intervals (bins) of 1 kb (for 400 bins/kb), and returns the position of that SNV.

helpers.py (Downstream)

```
import csv
import numpy
import pandas

# Search for matches, returning a series of Nulls and values matching
the condition.
def get_matches_in_column_in_range(df, col, lower_limit, upper_limit):
     potential_matches_vector = df[col].where((df[col] > lower_limit) &
(df[col] < upper_limit))

     # Trim the series of null values.
     potential_matches_vector =
potential_matches_vector[~potential_matches_vector.isnull()]

     return potential_matches_vector

# Get a series containing all matches
def get_first_match_in_column_in_range(df, col, lower_limit,
upper_limit):
     potential_matches_vector = get_matches_in_column_in_range(df, col,
lower_limit, upper_limit)

     # Return the first match, or Nil otherwise
     match = numpy.nan if len(potential_matches_vector) < 1 else
potential_matches_vector.iloc[0]
     index = numpy.nan if len(potential_matches_vector) < 1 else
potential_matches_vector.index[0]
     return index, match
```

---

inter peak coordination.py (Downstream)

```
import csv
import numpy as np
import pandas as pd
import helpers

# CSV file constants
INPUT_FILE_NAME = 'Input.csv'
OUTPUT_FILE_NAME = 'Ouptut.csv'
POSITION_1 = 'pos'
```

```
CHROMOSOME_1 = 'chr'
POSITION_2 = 'pos_2'
CHROMOSOME_2 = 'chr_2'

# Bin constants
BINS = 400
BIN_LIST = range(1, BINS+1)
INITIAL_LOWER_LIMIT = 1000
BIN_STEP_SIZE = 1000

# Read the inital data frame from the csv file, and
# append a column for each bin to the data frame
print('Started reading the dataframe...')
data_frame = pd.read_csv(INPUT_FILE_NAME)
for abin in BIN_LIST:
     data_frame[str(abin)] = np.nan
print('Finished reading the dataframe')


# Iterate through the data points
print('Started processing the dataframe')
data_point_count = len(data_frame)

for index, data_point in data_frame.iterrows():
     # Reset the limit values at each iteration, and
     # get the current position
     current_limit = INITIAL_LOWER_LIMIT
     pos = data_point[POSITION_1]

     # Iterate through the bins
     for abin in BIN_LIST:
       lower_limit = current_limit
       upper_limit = lower_limit + BIN_STEP_SIZE

       # Get the first match in the second position column, and
       # append the match to the dataframe
       match_index, match =
helpers.get_first_match_in_column_in_range(data_frame, POSITION_2, pos
+ lower_limit, pos + upper_limit)

       if match_index is not np.nan:
         matching_data_point_chromosome =
         data_frame.iloc[match_index][CHROMOSOME_2]
         if data_point[CHROMOSOME_1] !=
         matching_data_point_chromosome:
           data_frame.at[index, str(abin)] = 0
         else:
           data_frame.at[index, str(abin)] = match
       else:
         data_frame.at[index, str(abin)] = 0


       # Update the value of the limit
       current_limit = upper_limit
print('Finished processing the dataframe')
```

```
# Write the resulting dataframe to the output file
data_frame.to_csv(OUTPUT_FILE_NAME)
```

```
#Similar to the Python code above, this function anchors each cis SNV
and searches for other SNVs in upstream incremental genomic intervals
(bins) of 1 kb (for 400 bins/kb), and returns the position of that SNV.
```

helpers.py (Upstream)

```
import csv
import numpy
import pandas

# Search for matches, returning a series of Nulls and values matching
the condition.
def get_matches_in_column_in_range(df, col, lower_limit, upper_limit):
     potential_matches_vector = df[col].where((df[col] < lower_limit) &
(df[col] > upper_limit))

     # Trim the series of null values.
     potential_matches_vector =
potential_matches_vector[~potential_matches_vector.isnull()]

     return potential_matches_vector

# Get a series containing all matches
def get_first_match_in_column_in_range(df, col, lower_limit,
upper_limit):
     potential_matches_vector = get_matches_in_column_in_range(df, col,
lower_limit, upper_limit)

     # Return the first match, or Nil otherwise
     match = numpy.nan if len(potential_matches_vector) < 1 else
potential_matches_vector.iloc[0]
     index = numpy.nan if len(potential_matches_vector) < 1 else
potential_matches_vector.index[0]
     return index, match
```

---

inter peak coordination.py (Downstream)

```
import csv
import numpy as np
import pandas as pd
import helpers

# CSV file constants
INPUT_FILE_NAME = 'Input.csv'
OUTPUT_FILE_NAME = 'Ouptut.csv'
POSITION_1 = 'pos'
CHROMOSOME_1 = 'chr'
POSITION_2 = 'pos_2'
CHROMOSOME_2 = 'chr_2'
```

```
# Bin constants
BINS = 400
BIN_LIST = range(1, BINS+1)
INITIAL_LOWER_LIMIT = -1000
BIN_STEP_SIZE = 1000

# Read the inital data frame from the csv file, and
# append a column for each bin to the data frame
print('Started reading the dataframe...')
data_frame = pd.read_csv(INPUT_FILE_NAME)
for abin in BIN_LIST:
     data_frame[str(abin)] = np.nan
print('Finished reading the dataframe')


# Iterate through the data points
print('Started processing the dataframe')
data_point_count = len(data_frame)

for index, data_point in data_frame.iterrows():
     # Reset the limit values at each iteration, and
     # get the current position
     current_limit = INITIAL_LOWER_LIMIT
     pos = data_point[POSITION_1]

     # Iterate through the bins
     for abin in BIN_LIST:
        lower_limit = current_limit
        upper_limit = lower_limit - BIN_STEP_SIZE

        # Get the first match in the second position column, and
        # append the match to the dataframe
        match_index, match =
helpers.get_first_match_in_column_in_range(data_frame, POSITION_2, pos
+ lower_limit, pos + upper_limit)

        if match_index is not np.nan:
          matching_data_point_chromosome =
          data_frame.iloc[match_index][CHROMOSOME_2]
          if data_point[CHROMOSOME_1] !=
          matching_data_point_chromosome:
            data_frame.at[index, str(abin)] = 0
          else:
            data_frame.at[index, str(abin)] = match
        else:
          data_frame.at[index, str(abin)] = 0


        # Update the value of the limit
        current_limit = upper_limit
print('Finished processing the dataframe')

# Write the resulting dataframe to the output file
data_frame.to_csv(OUTPUT_FILE_NAME)
```

#The outputs of the two python codes above are combined so that all SNVs in up/downstream bins are now in one bin for the same incremental interval. The positions of the SNVs are then replaced by their F1 allelic ratios (BL6/(BL6+CAST)). Spearman's correlation coefficient of allelic ratios (BL6/(BL6+CAST)) is then computed between cis-acting variants and the SNVs in each successive bin. The following R code takes a dataframe of each cis and its F1 ratio and columns describing the F1 ratios for all SNVs in each bin. Spearman's ρ for each mutually exclusive bin with the corresponding anchor cis CTCF site is then calculated and used as the outcome variable in a linear regression model.#

inter peak Spearman correlation.R

```
corr_res_400_alt = c()
for (i in 4:ncol(corr_df_400bin)){
     corr = cbind(corr_df_400bin$f1_ratio,corr_df_400bin[,i])
     corr = as.data.frame(corr)
     corr = filter(corr, corr[,1] > 0, corr[,2] > 0,)
     sp_corr = cor.test(corr$V1, corr$V2,method = "spearman",
exact=FALSE)
     corr_res_400_alt = rbind(corr_res_400_alt, sp_corr$estimate)
}

corr_x_400_alt = 1:400
corr_res_400_alt = cbind(corr_x_400_alt,corr_res_400_alt[,1])
colnames(corr_res_400_alt) = c("Bin_kb", "rho")
corr_res_400_alt = as.data.frame(corr_res_400_alt)
```

References

1. Edwards SL, Beesley J, French JD, Dunning AM: **Beyond GWASs: illuminating the dark road from association to function.** *Am J Hum Genet* 2013, **93:**779-797.

2. Ward MC, Gilad Y: **Human genomics: Cracking the regulatory code.** *Nature* 2017, **550:**190-191.

3. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489:**75-82.

4. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al: **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).** *Nucleic Acids Res* 2017, **45:**D896-D901.

5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.

6. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

7. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.

8. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol* 2010, **28:**1045-1048.

9. e GP: **Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease.** *Nat Genet* 2017, **49:**1664-1670.

10. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome Res* 2011, **21:**1757-1767.

11. Lenhard B, Sandelin A, Carninci P: **Metazoan promoters: emerging characteristics and insights into transcriptional regulation.** *Nat Rev Genet* 2012, **13:**233-245.

12. Fuda NJ, Ardehali MB, Lis JT: **Defining mechanisms that regulate RNA polymerase II transcription in vivo.** *Nature* 2009, **461:**186-192.

13. Levine M: **Transcriptional enhancers in animal development and evolution.** *Curr Biol* 2010, **20:**R754-763.

14. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome.** *Annu Rev Genomics Hum Genet* 2006, **7:**29-59.

15. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G: **Enhancers: five essential questions.** *Nat Rev Genet* 2013, **14:**288-295.

16. Shlyueva D, Stampfel G, Stark A: **Transcriptional enhancers: from properties to genome-wide predictions.** *Nat Rev Genet* 2014, **15:**272-286.

17. Gibcus JH, Dekker J: **The hierarchy of the 3D genome.** *Mol Cell* 2013, **49:**773-782.

18. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459:**108-112.

19. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans.** *Nature* 2011, **470:**279-283.

20. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518:**317-330.

21. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448:**553-560.

22. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4:**651-657.

23. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nat Methods* 2009, **6:**283-289.

24. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods* 2013, **10:**1213-1218.

25. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, et al: **Ensembl 2017.** *Nucleic Acids Res* 2017, **45:**D635-D642.

26. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473:**337-342.

27. Arrowsmith CH, Bountra C, Fish PV, Lee K, Schapira M: **Epigenetic protein families: a new frontier for drug discovery.** *Nat Rev Drug Discov* 2012, **11:**384-400.

28. Dawson MA, Kouzarides T: **Cancer epigenetics: from mechanism to therapy.** *Cell* 2012, **150:**12-27.

29. Bulger M, Groudine M: **Functional and mechanistic diversity of distal transcription enhancers.** *Cell* 2011, **144:**327-339.

30. Petrykowska HM, Vockley CM, Elnitski L: **Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus.** *Genome Res* 2008, **18:**1238-1246.

31. Vokes SA, Ji H, Wong WH, McMahon AP: **A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb.** *Genes Dev* 2008, **22:**2651-2663.

32. Gaszner M, Felsenfeld G: **Insulators: exploiting transcriptional and epigenetic mechanisms.** *Nat Rev Genet* 2006, **7:**703-713.

33. Calhoun VC, Stathopoulos A, Levine M: **Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex.** *Proc Natl Acad Sci U S A* 2002, **99:**9243-9247.

34. Roeder RG: **The role of general initiation factors in transcription by RNA polymerase II.** *Trends Biochem Sci* 1996, **21:**327-335.

35. Spitz F, Furlong EE: **Transcription factors: from enhancer binding to developmental control.** *Nat Rev Genet* 2012, **13:**613-626.

36. Geyer PK, Green MM, Corces VG: **Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila.** *EMBO J* 1990, **9:**2247-2256.

37. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R: **Interchromosomal interactions and olfactory receptor choice.** *Cell* 2006, **126:**403-413.

38. Stamatoyannopoulos J: **Connecting the regulatory genome.** *Nat Genet* 2016, **48:**479-480.

39. Sanyal A, Lajoie BR, Jain G, Dekker J: **The long-range interaction landscape of gene promoters.** *Nature* 2012, **489:**109-113.

40. Clapier CR, Cairns BR: **The biology of chromatin remodeling complexes.** *Annu Rev Biochem* 2009, **78:**273-304.

41. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.

42. Struhl K: **Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast.** *Cell* 1987, **49:**295-297.

43. Meng H, Bartholomew B: **Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II.** *J Biol Chem* 2018, **293:**13786-13794.

44. Werner F: **Structural evolution of multisubunit RNA polymerases.** *Trends Microbiol* 2008, **16:**247-250.

45. Vannini A, Cramer P: **Conservation between the RNA polymerase I, II, and III transcription initiation machineries.** *Mol Cell* 2012, **45:**439-446.

46. Boeger H, Bushnell DA, Davis R, Griesenbeck J, Lorch Y, Strattan JS, Westover KD, Kornberg RD: **Structural basis of eukaryotic gene transcription.** *FEBS Lett* 2005, **579:**899-903.

47. Goodrich JA, Cutler G, Tjian R: **Contacts in context: promoter specificity and macromolecular interactions in transcription.** *Cell* 1996, **84:**825-830.

48. Fishburn J, Tomko E, Galburt E, Hahn S: **Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation.** *Proc Natl Acad Sci U S A* 2015, **112:**3961-3966.

49. He Y, Fang J, Taatjes DJ, Nogales E: **Structural visualization of key steps in human transcription initiation.** *Nature* 2013, **495:**481-486.

50. Grunberg S, Warfield L, Hahn S: **Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening.** *Nat Struct Mol Biol* 2012, **19:**788-796.

51. Kim TK, Ebright RH, Reinberg D: **Mechanism of ATP-dependent promoter melting by transcription factor IIH.** *Science* 2000, **288:**1418-1422.

52. Saunders A, Core LJ, Lis JT: **Breaking barriers to transcription elongation.** *Nat Rev Mol Cell Biol* 2006, **7:**557-567.

53. Venters BJ, Pugh BF: **A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome.** *Genome Res* 2009, **19:**360-371.

54. Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, Adelman K, Levine M, Young RA: **RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo.** *Nat Genet* 2007, **39:**1512-1516.

55. Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K: **RNA polymerase is poised for activation across the genome.** *Nat Genet* 2007, **39:**1507-1511.

56. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells.** *Cell* 2007, **130:**77-88.

57. Kornberg RD: **The molecular basis of eukaryotic transcription.** *Proc Natl Acad Sci U S A* 2007, **104:**12955-12961.

58. Peterson CL, Workman JL: **Promoter targeting and chromatin remodeling by the SWI/SNF complex.** *Curr Opin Genet Dev* 2000, **10:**187-192.

59. Larschan E, Winston F: **The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4.** *Genes Dev* 2001, **15:**1946-1956.

60. Ahn SH, Kim M, Buratowski S: **Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing.** *Mol Cell* 2004, **13:**67-76.

61. Stargell LA, Struhl K: **Mechanisms of transcriptional activation in vivo: two steps forward.** *Trends Genet* 1996, **12:**311-315.

62. Esnault C, Ghavi-Helm Y, Brun S, Soutourina J, Van Berkum N, Boschiero C, Holstege F, Werner M: **Mediator-dependent recruitment of TFIIH modules in preinitiation complex.** *Mol Cell* 2008, **31:**337-346.

63. Suzuki MM, Bird A: **DNA methylation landscapes: provocative insights from epigenomics.** *Nat Rev Genet* 2008, **9:**465-476.

64. Bird A: **The dinucleotide CG as a genomic signalling module.** *J Mol Biol* 2011, **409:**47-53.

65. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al: **Conserved role of intragenic DNA methylation in regulating alternative promoters.** *Nature* 2010, **466:**253-257.

66. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP: **Orphan CpG islands identify numerous conserved promoters in the mammalian genome.** *PLoS Genet* 2010, **6:**e1001134.

67. Kouzarides T: **Chromatin modifications and their function.** *Cell* 2007, **128:**693-705.

68. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.

69. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B: **Transcriptional features of genomic regulatory blocks.** *Genome Biol* 2009, **10:**R38.

70. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, Ohler U: **Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level.** *PLoS Genet* 2011, **7:**e1001274.

71. Yamashita R, Suzuki Y, Sugano S, Nakai K: **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity.** *Gene* 2005, **350:**129-136.

72. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B: **Genomic regulatory blocks underlie extensive microsynteny conservation in insects.** *Genome Res* 2007, **17:**1898-1908.

73. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST: **A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling.** *Cell* 2009, **138:**114-128.

74. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, et al: **CpG islands influence chromatin structure via the CpG-binding protein Cfp1.** *Nature* 2010, **464:**1082-1086.

75. Visel A, Rubin EM, Pennacchio LA: **Genomic views of distant-acting enhancers.** *Nature* 2009, **461:**199-205.

76. Kim TK, Shiekhattar R: **Architectural and Functional Commonalities between Enhancers and Promoters.** *Cell* 2015, **162:**948-959.

77. Akbari OS, Bae E, Johnsen H, Villaluz A, Wong D, Drewell RA: **A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the Drosophila embryo.** *Development* 2008, **135:**123-131.

78. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485:**381-385.

79. Gorkin DU, Leung D, Ren B: **The 3D genome in transcriptional regulation and pluripotency.** *Cell Stem Cell* 2014, **14:**762-775.

80. Ong CT, Corces VG: **CTCF: an architectural protein bridging genome topology and function.** *Nat Rev Genet* 2014, **15:**234-246.

81. Malik S, Roeder RG: **The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation.** *Nat Rev Genet* 2010, **11:**761-772.

82. Carvajal JJ, Cox D, Summerbell D, Rigby PW: **A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and maintenance of gene expression during muscle development.** *Development* 2001, **128:**1857-1868.

83. Spitz F, Gonzalez F, Duboule D: **A global control region defines a chromosomal regulatory landscape containing the HoxD cluster.** *Cell* 2003, **113:**405-417.

84. Marinic M, Aktas T, Ruf S, Spitz F: **An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape.** *Dev Cell* 2013, **24:**530-542.

85. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, Spitz F: **Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor.** *Nat Genet* 2011, **43:**379-386.

86. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:**499-502.

87. Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T: **Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb.** *Development* 2005, **132:**797-803.

88. Jeong Y, El-Jaick K, Roessler E, Muenke M, Epstein DJ: **A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers.** *Development* 2006, **133:**761-772.

89. Zippo A, Serafini R, Rocchigiani M, Pennacchini S, Krepelova A, Oliviero S: **Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation.** *Cell* 2009, **138:**1122-1136.

90. Gillies SD, Morrison SL, Oi VT, Tonegawa S: **A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene.** *Cell* 1983, **33:**717-728.

91. Banerji J, Olson L, Schaffner W: **A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes.** *Cell* 1983, **33:**729-740.

92. Mohibullah N, Hahn S: **Site-specific cross-linking of TBP in vivo and in vitro reveals a direct functional interaction with the SAGA subunit Spt3.** *Genes Dev* 2008, **22:**2994-3006.

93. Huisinga KL, Pugh BF: **A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae.** *Mol Cell* 2004, **13:**573-585.

94. Tirosh I, Barkai N: **Two strategies for gene regulation by promoter nucleosomes.** *Genome Res* 2008, **18:**1084-1091.

95. Roux BT, Lindsay MA, Heward JA: **Knockdown of Nuclear-Located Enhancer RNAs and Long ncRNAs Using Locked Nucleic Acid GapmeRs.** *Methods Mol Biol* 2017, **1468:**11-18.

96. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL: **A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells.** *Cell* 2011, **145:**622-634.

97. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.** *Nat Rev Genet* 2013, **14:**390-403.

98. Lee TI, Young RA: **Transcriptional regulation and its misregulation in disease.** *Cell* 2013, **152:**1237-1251.

99. Ong CT, Corces VG: **Enhancer function: new insights into the regulation of tissue-specific gene expression.** *Nat Rev Genet* 2011, **12:**283-293.

100. Makova KD, Hardison RC: **The effects of chromatin organization on variation in mutation rates in the genome.** *Nat Rev Genet* 2015, **16:**213-223.

101. Levine M, Cattoglio C, Tjian R: **Looping back to leap forward: transcription enters a new era.** *Cell* 2014, **157:**13-25.

102. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS, et al: **Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells.** *Nat Genet* 2010, **42:**53-61.

103. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA: **Master transcription factors and mediator establish super-enhancers at key cell identity genes.** *Cell* 2013, **153:**307-319.

104. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA: **Super-enhancers in the control of cell identity and disease.** *Cell* 2013, **155:**934-947.

105. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA: **Selective inhibition of tumor oncogenes by disruption of super-enhancers.** *Cell* 2013, **153:**320-334.

106. Ing-Simmons E, Seitan VC, Faure AJ, Flicek P, Carroll T, Dekker J, Fisher AG, Lenhard B, Merkenschlager M: **Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin.** *Genome Res* 2015, **25:**504-513.

107. Siersbaek R, Rabiee A, Nielsen R, Sidoli S, Traynor S, Loft A, Poulsen LC, Rogowska-Wrzesinska A, Jensen ON, Mandrup S: **Transcription factor cooperativity in early adipogenic hotspots and super-enhancers.** *Cell Rep* 2014, **7:**1443-1455.

108. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SC, Erdos MR, Davis SR, Roychoudhuri R, Restifo NP, Gadina M, et al: **Super-enhancers delineate disease-associated regulatory nodes in T cells.** *Nature* 2015, **520:**558-562.

109. Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon KR, Resch W, Liang G, Tang Z, Mathe E, Benner C, et al: **B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity.** *Cell* 2014, **159:**1524-1537.

110. Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar AM, Sharpe JA, Suciu MC, et al: **Genetic dissection of the alpha-globin super-enhancer in vivo.** *Nat Genet* 2016, **48:**895-903.

111. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G: **A large fraction of extragenic RNA pol II transcription sites overlap enhancers.** *PLoS Biol* 2010, **8:**e1000384.

112. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al: **Widespread transcription at neuronal activity-regulated enhancers.** *Nature* 2010, **465:**182-187.

113. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101-108.

114. Natoli G, Andrau JC: **Noncoding transcription at enhancers: general principles and functional models.** *Annu Rev Genet* 2012, **46:**1-19.

115. Struhl K: **Transcriptional noise and the fidelity of initiation by RNA polymerase II.** *Nat Struct Mol Biol* 2007, **14:**103-105.

116. Wu H, Nord AS, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA, Visel A: **Tissue-specific RNA expression marks distant-acting developmental enhancers.** *PLoS Genet* 2014, **10:**e1004610.

117. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: **An atlas of active enhancers across human cell types and tissues.** *Nature* 2014, **507:**455-461.

118. Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu ZP, et al: **RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity.** *Cell* 2015, **161:**774-789.

119. Aguilo F, Li S, Balasubramaniyan N, Sancho A, Benko S, Zhang F, Vashisht A, Rengasamy M, Andino B, Chen CH, et al: **Deposition of 5-Methylcytosine on Enhancer RNAs Enables the Coactivator Function of PGC-1alpha.** *Cell Rep* 2016, **14:**479-492.

120. Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP: **A comprehensive nuclear receptor network for breast cancer cells.** *Cell Rep* 2013, **3:**538-551.

121. Liu Z, Merkurjev D, Yang F, Li W, Oh S, Friedman MJ, Song X, Zhang F, Ma Q, Ohgi KA, et al: **Enhancer activation requires trans-recruitment of a mega transcription factor complex.** *Cell* 2014, **159:**358-373.

122. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, Lieb JD, Widom J, Segal E, Hughes TR: **High nucleosome occupancy is encoded at human regulatory sequences.** *PLoS One* 2010, **5:**e9129.

123. Dubois-Chevalier J, Dubois V, Dehondt H, Mazrooei P, Mazuy C, Serandour AA, Gheeraert C, Guillaume P, Bauge E, Derudas B, et al: **The logic of transcriptional regulator recruitment architecture at cis-regulatory modules controlling liver functions.** *Genome Res* 2017, **27:**985-996.

124. Mirny LA: **Nucleosome-mediated cooperativity between transcription factors.** *Proc Natl Acad Sci U S A* 2010, **107:**22534-22539.

125. Miller JA, Widom J: **Collaborative competition mechanism for gene activation in vivo.** *Mol Cell Biol* 2003, **23:**1623-1632.

126. Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N: **Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model.** *Nat Genet* 2013, **45:**1021-1028.

127. Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EE: **Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity.** *PLoS Genet* 2014, **10:**e1004060.

128. Zaret KS, Carroll JS: **Pioneer transcription factors: establishing competence for gene expression.** *Genes Dev* 2011, **25:**2227-2241.

129. Zaret KS, Mango SE: **Pioneer transcription factors, chromatin dynamics, and cell fate control.** *Curr Opin Genet Dev* 2016, **37:**76-81.

130. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M: **FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription.** *Cell* 2008, **132:**958-970.

131. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS: **FOXA1 is a key determinant of estrogen receptor function and endocrine response.** *Nat Genet* 2011, **43:**27-33.

132. Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, Karpova TS, Ball D, Mazza D, Lavis LD, et al: **Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions.** *Cell* 2016, **165:**593-605.

133. Liu Z, Kraus WL: **Catalytic-Independent Functions of PARP-1 Determine Sox2 Pioneer Activity at Intractable Genomic Loci.** *Mol Cell* 2017, **65:**589-603 e589.

134. Chen J, Zhang Z, Li L, Chen BC, Revyakin A, Hajj B, Legant W, Dahan M, Lionnet T, Betzig E, et al: **Single-molecule dynamics of enhanceosome assembly in embryonic stem cells.** *Cell* 2014, **156:**1274-1285.

135. Franco HL, Nagari A, Kraus WL: **TNFalpha signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome.** *Mol Cell* 2015, **58:**21-34.

136. Soufi A, Donahue G, Zaret KS: **Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome.** *Cell* 2012, **151:**994-1004.

137. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS: **Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming.** *Cell* 2015, **161:**555-568.

138. Donaghey J, Thakurela S, Charlton J, Chen JS, Smith ZD, Gu H, Pop R, Clement K, Stamenova EK, Karnik R, et al: **Genetic determinants and epigenetic effects of pioneer-factor occupancy.** *Nat Genet* 2018, **50:**250-258.

139. Taatjes DJ, Marr MT, Tjian R: **Regulatory diversity among metazoan co-activator complexes.** *Nat Rev Mol Cell Biol* 2004, **5:**403-410.

140. Krasnov AN, Mazina MY, Nikolenko JV, Vorobyeva NE: **On the way of revealing coactivator complexes cross-talk during transcriptional activation.** *Cell Biosci* 2016, **6:**15.

141. Sakabe NJ, Nobrega MA: **Beyond the ENCODE project: using genomics and epigenomics strategies to study enhancer evolution.** *Philos Trans R Soc Lond B Biol Sci* 2013, **368:**20130022.

142. Calo E, Wysocka J: **Modification of enhancer chromatin: what, how, and why?** *Mol Cell* 2013, **49:**825-837.

143. Whitaker JW, Nguyen TT, Zhu Y, Wildberg A, Wang W: **Computational schemes for the prediction and annotation of enhancers from epigenomic assays.** *Methods* 2015, **72:**86-94.

144. Blackwood EM, Kadonaga JT: **Going the distance: a current view of enhancer action.** *Science* 1998, **281:**60-63.

145. Grosveld F, van Assendelft GB, Greaves DR, Kollias G: **Position-independent, high-level expression of the human beta-globin gene in transgenic mice.** *Cell* 1987, **51:**975-985.

146. Vernimmen D, Bickmore WA: **The Hierarchy of Transcriptional Activation: From Enhancer to Promoter.** *Trends Genet* 2015, **31:**696-708.

147. Hsieh CL, Fei T, Chen Y, Li T, Gao Y, Wang X, Sun T, Sweeney CJ, Lee GS, Chen S, et al: **Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation.** *Proc Natl Acad Sci U S A* 2014, **111:**7319-7324.

148. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R: **Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.** *Nature* 2013, **494:**497-501.

149. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34:**D95-97.

150. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.

151. Vingron M, Brazma A, Coulson R, van Helden J, Manke T, Palin K, Sand O, Ukkonen E: **Integrating sequence, evolution and functional genomics in regulatory genomics.** *Genome Biol* 2009, **10:**202.

152. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, et al: **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.** *Science* 2013, **342:**744-747.

153. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, Guigo R, Iossifov I, Vasileva A, Lappalainen T: **Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk.** *Nat Genet* 2018, **50:**1327-1334.

154. Klann TS, Black JB, Chellappan M, Safi A, Song L, Hilton IB, Crawford GE, Reddy TE, Gersbach CA: **CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome.** *Nat Biotechnol* 2017, **35:**561-568.

155. Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, Martchenko A, So V, Macpherson NN, Mitchell JA: **Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes.** *Genome Res* 2017, **27:**246-258.

156. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, et al: **Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element.** *Science* 2014, **346:**1373-1377.

157. Abraham BJ, Hnisz D, Weintraub AS, Kwiatkowski N, Li CH, Li Z, Weichert-Leahey N, Rahman S, Liu Y, Etchin J, et al: **Small genomic insertions form enhancers that misregulate oncogenes.** *Nat Commun* 2017, **8:**14385.

158. Barolo S: **Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy.** *Bioessays* 2012, **34:**135-141.

159. Miguel-Escalada I, Pasquali L, Ferrer J: **Transcriptional enhancers: functional insights and role in human disease.** *Curr Opin Genet Dev* 2015, **33:**71-76.

160. Strickfaden H, Zunhammer A, van Koningsbruggen S, Kohler D, Cremer T: **4D chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited.** *Nucleus* 2010, **1:**284-297.

161. Orlova DY, Stixova L, Kozubek S, Gierman HJ, Sustackova G, Chernyshev AV, Medvedev RN, Legartova S, Versteeg R, Matula P, et al: **Arrangement of nuclear structures is not transmitted through mitosis but is identical in sister cells.** *J Cell Biochem* 2012, **113:**3313-3329.

162. Parada LA, Roix JJ, Misteli T: **An uncertainty principle in chromosome positioning.** *Trends Cell Biol* 2003, **13:**393-396.

163. Lieb JD, Clarke ND: **Control of transcription through intragenic patterns of nucleosome composition.** *Cell* 2005, **123:**1187-1190.

164. Rando OJ, Ahmad K: **Rules and regulation in the primary structure of chromatin.** *Curr Opin Cell Biol* 2007, **19:**250-256.

165. Strahl BD, Allis CD: **The language of covalent histone modifications.** *Nature* 2000, **403:**41-45.

166. Sauvageau M, Sauvageau G: **Polycomb group proteins: multi-faceted regulators of somatic stem cells and cancer.** *Cell Stem Cell* 2010, **7:**299-313.

167. Schuettengruber B, Cavalli G: **Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice.** *Development* 2009, **136:**3531-3542.

168. Sawarkar R, Paro R: **Interpretation of developmental signaling at chromatin: the Polycomb perspective.** *Dev Cell* 2010, **19:**651-661.

169. Spielmann M, Lupianez DG, Mundlos S: **Structural variation in the 3D genome.** *Nat Rev Genet* 2018, **19:**453-467.

170. Phillips JE, Corces VG: **CTCF: master weaver of the genome.** *Cell* 2009, **137:**1194-1211.

171. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L: **Genome architectures revealed by tethered chromosome conformation capture and population-based modeling.** *Nat Biotechnol* 2011, **30:**90-98.

172. Bohn M, Heermann DW: **Diffusion-driven looping provides a consistent framework for chromatin organization.** *PLoS One* 2010, **5:**e12218.

173. van Steensel B, Dekker J: **Genomics tools for unraveling chromosome architecture.** *Nat Biotechnol* 2010, **28:**1089-1095.

174. Bau D, Marti-Renom MA: **Structure determination of genomic domains by satisfaction of spatial restraints.** *Chromosome Res* 2011, **19:**25-35.

175. Hakim O, Misteli T: **SnapShot: Chromosome confirmation capture.** *Cell* 2012, **148:**1068 e1061-1062.

176. Fudenberg G, Mirny LA: **Higher-order chromatin structure: bridging physics and biology.** *Curr Opin Genet Dev* 2012, **22:**115-124.

177. Felsenfeld G, Groudine M: **Controlling the double helix.** *Nature* 2003, **421:**448-453.

178. Henikoff S, Shilatifard A: **Histone modification: cause or cog?** *Trends Genet* 2011, **27:**389-396.

179. Martin C, Zhang Y: **Mechanisms of epigenetic inheritance.** *Curr Opin Cell Biol* 2007, **19:**266-272.

180. Ruthenburg AJ, Allis CD, Wysocka J: **Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark.** *Mol Cell* 2007, **25:**15-30.

181. Bajpai R, Chen DA, Rada-Iglesias A, Zhang J, Xiong Y, Helms J, Chang CP, Zhao Y, Swigut T, Wysocka J: **CHD7 cooperates with PBAF to control multipotent neural crest formation.** *Nature* 2010, **463:**958-962.

182. Hargreaves DC, Crabtree GR: **ATP-dependent chromatin remodeling: genetics, genomics and mechanisms.** *Cell Res* 2011, **21:**396-420.

183. Waterborg JH: **Dynamics of histone acetylation in vivo. A function for acetylation turnover?** *Biochem Cell Biol* 2002, **80:**363-378.

184. Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, Peterson CL: **Histone H4-K16 acetylation controls chromatin structure and protein interactions.** *Science* 2006, **311:**844-847.

185. Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, Khorasanizadeh S: **Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains.** *Genes Dev* 2003, **17:**1870-1881.

186. Cao R, Zhang Y: **The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3.** *Curr Opin Genet Dev* 2004, **14:**155-164.

187. Smith E, Lin C, Shilatifard A: **The super elongation complex (SEC) and MLL in development and disease.** *Genes Dev* 2011, **25:**661-672.

188. Bungard D, Fuerth BJ, Zeng PY, Faubert B, Maas NL, Viollet B, Carling D, Thompson CB, Jones RG, Berger SL: **Signaling kinase AMPK activates stress-promoted transcription via histone H2B phosphorylation.** *Science* 2010, **329:**1201-1205.

189. Turner BM: **Environmental sensing by chromatin: an epigenetic contribution to evolutionary change.** *FEBS Lett* 2011, **585:**2032-2040.

190. Gardner KE, Allis CD, Strahl BD: **Operating on chromatin, a colorful language where context matters.** *J Mol Biol* 2011, **409:**36-46.

191. Hublitz P, Albert M, Peters AH: **Mechanisms of transcriptional repression by histone lysine methylation.** *Int J Dev Biol* 2009, **53:**335-354.

192. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, et al: **Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification.** *Cell* 2011, **146:**1016-1028.

193. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-837.

194. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40:**897-903.

195. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nat Genet* 2007, **39:**311-318.

196. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes.** *Nat Rev Genet* 2011, **12:**7-18.

197. Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, et al: **Distinct factors control histone variant H3.3 localization at specific genomic regions.** *Cell* 2010, **140:**678-691.

198. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nat Biotechnol* 2010, **28:**817-825.

199. Hon G, Wang W, Ren B: **Discovery and annotation of functional chromatin signatures in the human genome.** *PLoS Comput Biol* 2009, **5:**e1000566.

200. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D: **Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.** *Nat Genet* 2007, **39:**457-466.

201. Simon JA, Kingston RE: **Mechanisms of polycomb gene silencing: knowns and unknowns.** *Nat Rev Mol Cell Biol* 2009, **10:**697-708.

202. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung WK, Shahab A, Kuznetsov VA, et al: **Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells.** *Cell Stem Cell* 2007, **1:**286-298.

203. Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R, Thomson JA: **Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells.** *Cell Stem Cell* 2007, **1:**299-312.

204. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al: **Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.** *PLoS Genet* 2008, **4:**e1000242.

205. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125:**315-326.

206. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D: **Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors.** *Mol Cell* 2008, **30:**755-766.

207. Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, Brockdorff N, Fisher AG, Pombo A: **Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells.** *Nat Cell Biol* 2007, **9:**1428-1435.

208. Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA: **Enhancer decommissioning by LSD1 during embryonic stem cell differentiation.** *Nature* 2012, **482:**221-225.

209. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *Proc Natl Acad Sci U S A* 2010, **107:**21931-21936.

210. Rivera CM, Ren B: **Mapping human epigenomes.** *Cell* 2013, **155:**39-55.

211. Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, Glass CK: **Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription.** *Mol Cell* 2013, **51:**310-325.

212. Buecker C, Wysocka J: **Enhancers as information integration hubs in development: lessons from genomics.** *Trends Genet* 2012, **28:**276-284.

213. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43-49.

214. Wang A, Yue F, Li Y, Xie R, Harper T, Patel NA, Muth K, Palmer J, Qiu Y, Wang J, et al: **Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates.** *Cell Stem Cell* 2015, **16:**386-399.

215. Rada-Iglesias A, Wysocka J: **Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease.** *Genome Med* 2011, **3:**36.

216. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA: **Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor.** *Cell* 2012, **149:**1233-1244.

217. Dekker J: **Gene regulation in the third dimension.** *Science* 2008, **319:**1793-1794.

218. Kadauke S, Blobel GA: **Chromatin loops in gene regulation.** *Biochim Biophys Acta* 2009, **1789:**17-25.

219. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159:**1665-1680.

220. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, Cavalli G: **Multiscale 3D Genome Rewiring during Mouse Neural Development.** *Cell* 2017, **171:**557-572 e524.

221. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW, et al: **The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements.** *Genome Res* 2015, **25:**582-597.

222. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Varnai C, Thiecke MJ, et al: **Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.** *Cell* 2016, **167:**1369-1384 e1319.

223. Freire-Pritchett P, Schoenfelder S, Varnai C, Wingett SW, Cairns J, Collier AJ, Garcia-Vilchez R, Furlan-Magaril M, Osborne CS, Fraser P, et al: **Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells.** *Elife* 2017, **6**.

224. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al: **Mediator and cohesin connect gene expression and chromatin architecture.** *Nature* 2010, **467:**430-435.

225. Tan-Wong SM, Zaugg JB, Camblong J, Xu Z, Zhang DW, Mischo HE, Ansari AZ, Luscombe NM, Steinmetz LM, Proudfoot NJ: **Gene loops enhance transcriptional directionality.** *Science* 2012, **338:**671-675.

226. Yan J, Chen SA, Local A, Liu T, Qiu Y, Dorighi KM, Preissl S, Rivera CM, Wang C, Ye Z, et al: **Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers.** *Cell Res* 2018, **28:**204-220.

227. Dixon JR, Gorkin DU, Ren B: **Chromatin Domains: The Unit of Chromosome Organization.** *Mol Cell* 2016, **62:**668-680.

228. Dekker J, Mirny L: **The 3D Genome as Moderator of Chromosomal Communication.** *Cell* 2016, **164:**1110-1121.

229. Misteli T: **Beyond the sequence: cellular organization of genome function.** *Cell* 2007, **128:**787-800.

230. Fraser P, Bickmore W: **Nuclear organization of the genome and the potential for gene regulation.** *Nature* 2007, **447:**413-417.

231. Naumova N, Dekker J: **Integrating one-dimensional and three-dimensional maps of genomes.** *J Cell Sci* 2010, **123:**1979-1988.

232. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289-293.

233. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DC, Aitken S, et al: **Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation.** *Mol Syst Biol* 2015, **11:**852.

234. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y, et al: **Architectural protein subclasses shape 3D organization of genomes during lineage commitment.** *Cell* 2013, **153:**1281-1295.

235. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376-380.

236. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148:**458-472.

237. Hubner MR, Eckersley-Maslin MA, Spector DL: **Chromatin organization and transcriptional regulation.** *Curr Opin Genet Dev* 2013, **23:**89-95.

238. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture.** *Cell Rep* 2015, **10:**1297-1309.

239. Cubenas-Potts C, Corces VG: **Topologically Associating Domains: An invariant framework or a dynamic scaffold?** *Nucleus* 2015, **6:**430-434.

240. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, Cubenas-Potts C, Hu M, Lei EP, Bosco G, et al: **Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing.** *Mol Cell* 2015, **58:**216-231.

241. Narendra V, Bulajic M, Dekker J, Mazzoni EO, Reinberg D: **CTCF-mediated topological boundaries during development foster appropriate gene regulation.** *Genes Dev* 2016, **30:**2657-2662.

242. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B: **A map of the cis-regulatory sequences in the mouse genome.** *Nature* 2012, **488:**116-120.

243. Zhan Y, Mariani L, Barozzi I, Schulz EG, Bluthgen N, Stadler M, Tiana G, Giorgetti L: **Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes.** *Genome Res* 2017, **27:**479-490.

244. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al: **Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.** *Cell* 2015, **161:**1012-1025.

245. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al: **Activation of proto-oncogenes by disruption of chromosome neighborhoods.** *Science* 2016, **351:**1454-1458.

246. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, Aifantis I, Tsirigos A: **Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries.** *Nat Commun* 2018, **9:**542.

247. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, Fonseca NA, Huber W, C HH, Mirny L, Spitz F: **Two independent modes of chromatin organization revealed by cohesin removal.** *Nature* 2017, **551:**51-56.

248. Bianco S, Lupianez DG, Chiariello AM, Annunziatella C, Kraft K, Schopflin R, Wittler L, Andrey G, Vingron M, Pombo A, et al: **Polymer physics predicts the effects of structural variants on chromatin architecture.** *Nat Genet* 2018, **50:**662-667.

249. Mawhinney MT, Liu R, Lu F, Maksimoska J, Damico K, Marmorstein R, Lieberman PM, Urbanc B: **CTCF-Induced Circular DNA Complexes Observed by Atomic Force Microscopy.** *J Mol Biol* 2018, **430:**759-776.

250. Goldman RD, Gruenbaum Y, Moir RD, Shumaker DK, Spann TP: **Nuclear lamins: building blocks of nuclear architecture.** *Genes Dev* 2002, **16:**533-547.

251. Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Peterfia B, Solovei I, Cremer T, Dopazo J, Langst G: **Initial genomics of the human nucleolus.** *PLoS Genet* 2010, **6:**e1000889.

252. Schaefke B, Sun W, Li YS, Fang L, Chen W: **The evolution of posttranscriptional regulation.** *Wiley Interdiscip Rev RNA* 2018**:**e1485.

253. Biggin MD: **Animal Transcription Networks as Highly Connected, Quantitative Continua.** *Developmental Cell* 2011, **21:**611-626.

254. Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, Alsawadi A, Valenti P, Plaza S, Payre F, et al: **Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness.** *Cell* 2015, **160:**191-203.

255. Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS: **Suboptimization of developmental enhancers.** *Science* 2015, **350:**325-328.

256. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J: **DNA-dependent formation of transcription factor pairs alters their binding specificity.** *Nature* 2015, **527:**384-+.

257. Wong ES, Schmitt BM, Kazachenka A, Thybert D, Redmond A, Connor F, Rayner TF, Feig C, Ferguson-Smith AC, Marioni JC, et al: **Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution.** *Nat Commun* 2017, **8:**1092.

258. Chen ZJ, Birchler JA: *Polyploid and hybrid genomics.* Ames, Iowa: Wiley-Blackwell; 2013.

259. Gao Q, Sun W, Ballegeer M, Libert C, Chen W: **Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing.** *Mol Syst Biol* 2015, **11:**816.

260. Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC: **Extensive compensatory cis-trans regulation in the evolution of mouse gene expression.** *Genome Res* 2012, **22:**2376-2384.

261. McManus CJ, Coolon JD, Eipper-Mains J, Wittkopp PJ, Graveley BR: **Evolution of splicing regulatory networks in Drosophila.** *Genome Research* 2014, **24:**786-796.

262. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al: **Chromatin architecture reorganization during stem cell differentiation.** *Nature* 2015, **518:**331-336.

263. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.** *Cell* 2014, **159:**1665-1680.

264. Roy SS, Mukherjee AK, Chowdhury S: **Insights about genome function from spatial organization of the genome.** *Hum Genomics* 2018, **12:**8.

265. Martinez SR, Miranda JL: **CTCF terminal segments are unstructured.** *Protein Sci* 2010, **19:**1110-1116.

266. Klenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM, Goodwin GH, Neiman PE, Lobanenkov VV: **CTCF, A CONSERVED NUCLEAR FACTOR REQUIRED FOR OPTIMAL TRANSCRIPTIONAL ACTIVITY OF THE CHICKEN C-MYC GENE, IS AN 11-ZN-FINGER PROTEIN DIFFERENTIALLY EXPRESSED IN MULTIPLE FORMS.** *Molecular and Cellular Biology* 1993, **13:**7612-7624.

267. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al: **A high-resolution map of human evolutionary constraint using 29 mammals.** *Nature* 2011, **478:**476-482.

268. Fedoriw AM, Stein P, Svoboda P, Schultz RM, Bartolomei MS: **Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting.** *Science* 2004, **303:**238-240.

269. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT: **Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.** *Cell* 2012, **148:**335-348.

270. Hou C, Zhao H, Tanimoto K, Dean A: **CTCF-dependent enhancer-blocking by alternative chromatin loop formation.** *Proc Natl Acad Sci U S A* 2008, **105:**20398-20403.

271. Mateos-Langerak J, Bohn M, de Leeuw W, Giromus O, Manders EM, Verschure PJ, Indemans MH, Gierman HJ, Heermann DW, van Driel R, Goetze S: **Spatially confined folding of chromatin in the interphase nucleus.** *Proc Natl Acad Sci U S A* 2009, **106:**3812-3817.

272. Burcin M, Arnold R, Lutz M, Kaiser B, Runge D, Lottspeich F, Filippova GN, Lobanenkov VV, Renkawitz R: **Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF.** *Molecular and Cellular Biology* 1997, **17:**1281-1288.

273. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T: **The chromatin insulator CTCF and the emergence of metazoan diversity.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109:**17507-17512.

274. Sleutels F, Soochit W, Bartkuhn M, Heath H, Dienstbach S, Bergmaier P, Franke V, Rosa-Garrido M, van de Nobelen S, Caesar L, et al: **The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner.** *Epigenetics & Chromatin* 2012, **5**.

275. Soshnikova N, Montavon T, Leleu M, Galjart N, Duboule D: **Functional analysis of CTCF during mammalian limb development.** *Dev Cell* 2010, **19:**819-830.

276. Wan L-B, Pan H, Hannenhalli S, Cheng Y, Ma J, Fedoriw A, Lobanenkov V, Latham KE, Schultz RM, Bartolomei MS: **Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development.** *Development* 2008, **135:**2729-2738.

277. Beygo J, Citro V, Sparago A, De Crescenzo A, Cerrato F, Heitmann M, Rademacher K, Guala A, Enklaar T, Anichini C, et al: **The molecular function and clinical phenotype of partial deletions of the IGF2/H19 imprinting control region depends on the spatial arrangement of the remaining CTCF-binding sites.** *Human Molecular Genetics* 2013, **22:**544-557.

278. de Wit E, Bouwman BAM, Zhu Y, Klous P, Splinter E, Verstegen MJAM, Krijger PHL, Festuccia N, Nora EP, Welling M, et al: **The pluripotent genome in three dimensions is shaped around pluripotency factors.** *Nature* 2013, **501:**227-+.

279. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, Splinter E, Wijchers PJ, Krijger PHL, de laat W: **CTCF Binding Polarity Determines Chromatin Looping.** *Molecular Cell* 2015, **60:**676-684.

280. Dowen JM, Bilodeau S, Orlando DA, Huebner MR, Abraham BJ, Spector DL, Young RA: **Multiple Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements.** *Stem Cell Reports* 2013, **1:**371-378.

281. Guo C, Yoon HS, Franklin A, Jain S, Ebert A, Cheng H-L, Hansen E, Despo O, Bossen C, Vettermann C, et al: **CTCF-binding elements mediate control of V(D)J recombination.** *Nature* 2011, **477:**424-U182.

282. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al: **CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function.** *Cell* 2015, **162:**900-910.

283. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Hom D, Kayserili H, Opitz JM, Laxova R, et al: **Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.** *Cell* 2015, **161:**1012-1025.

284. Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, Reinberg D: **CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation.** *Science* 2015, **347:**1017-1021.

285. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al: **Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112:**E6456-E6465.

286. Shrimali S, Srivastava S, Varma G, Grinberg A, Pfeifer K, Srivastava M: **An ectopic CTCF-dependent transcriptional insulator influences the choice of V beta gene segments for VDJ recombination at TCR beta locus.** *Nucleic Acids Research* 2012, **40:**7753-7765.

287. Saldana-Meyer R, Gonzalez-Buendia E, Guerrero G, Narendra V, Bonasio R, Recillas-Targa F, Reinberg D: **CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53.** *Genes & Development* 2014, **28:**723-734.

288. Kung JT, Kesner B, An JY, Ahn JY, Cifuentes-Rojas C, Colognori D, Jeon Y, Szanto A, del Rosario BC, Pinter SF, et al: **Locus-Specific Targeting to the X Chromosome Revealed by the RNA Interactome of CTCF.** *Molecular Cell* 2015, **57:**361-375.

289. Bell AC, Felsenfeld G: **Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene.** *Nature* 2000, **405:**482-485.

290. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schuebeler D: **Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions.** *Plos Genetics* 2013, **9**.

291. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM: **CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus.** *Nature* 2000, **405:**486-489.

292. Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi CF, Wolffe A, Ohlsson R, Lobanenkov VV: **Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive.** *Current Biology* 2000, **10:**853-856.

293. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al: **Widespread plasticity in CTCF occupancy linked to DNA methylation.** *Genome Research* 2012, **22:**1680-1688.

294. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suva ML, Bernstein BE: **Insulator dysfunction and oncogene activation in IDH mutant gliomas.** *Nature* 2016, **529:**110-114.

295. Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K: **Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.** *Genome Research* 2009, **19:**24-32.

296. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, Renkawitz R, Russell S, White R: **CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex.** *Plos Genetics* 2007, **3:**1211-1222.

297. Nakahashi H, Kwon K-RK, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S, Yamane A, et al: **A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code.** *Cell Reports* 2013, **3:**1678-1689.

298. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128:**1231-1245.

299. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.** *Proc Natl Acad Sci U S A* 2007, **104:**7145-7150.

300. Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G, Pedone PV: **Critical DNA binding interactions of the insulator protein CTCF - A small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci.** *Journal of Biological Chemistry* 2007, **282:**33336-33345.

301. Rhee HS, Pugh BF: **Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution.** *Cell* 2011, **147:**1408-1419.

302. Xiao T, Wongtrakoongate P, Trainor C, Felsenfeld G: **CTCF Recruits Centromeric Protein CENP-E to the Pericentromeric/Centromeric Regions of Chromosomes through Unusual CTCF-Binding Sites.** *Cell Reports* 2015, **12:**1704-1714.

303. Ghirlando R, Felsenfeld G: **CTCF: making the right connections.** *Genes Dev* 2016, **30:**881-891.

304. Engel N, West AG, Felsenfeld G, Bartolomei MS: **Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations.** *Nature Genetics* 2004, **36:**883-888.

305. Rodriguez C, Borgel J, Court F, Cathala G, Forne T, Piette J: **CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus.** *Biochemical and Biophysical Research Communications* 2010, **392:**129-134.

306. Lai AY, Fatemi M, Dhasarathy A, Malone C, Sobol SE, Geigerman C, Jaye DL, Mav D, Shah R, Li L, Wade PA: **DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas.** *Journal of Experimental Medicine* 2010, **207:**1939-1950.

307. Chang J, Zhang B, Heath H, Galjart N, Wang X, Milbrandt J: **Nicotinamide adenine dinucleotide (NAD)-regulated DNA methylation alters CCCTC-binding factor (CTCF)/cohesin binding and transcription at the BDNF locus.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107:**21836-21841.

308. Zampieri M, Guastafierro T, Calabrese R, Ciccarone F, Bacalini MG, Reale A, Perilli M, Passananti C, Caiafa P: **ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites.** *Biochemical Journal* 2012, **441:**645-652.

309. Guastafierro T, Cecchinelli B, Zampieri M, Reale A, Riggio G, Sthandier O, Zupi G, Calabrese L, Caiafa P: **CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery.** *Journal of Biological Chemistry* 2008, **283:**21873-21880.

310. Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G: **Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA.** *Genes & Development* 2010, **24:**2543-2555.

311. Sun S, Del Rosario BC, Szanto A, Ogawa Y, Jeon Y, Lee JT: **Jpx RNA Activates Xist by Evicting CTCF.** *Cell* 2013, **153:**1537-1551.

312. Rudan MV, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture.** *Cell Reports* 2015, **10:**1297-1309.

313. Gomez-Marin C, Tena JJ, Acemel RD, Lopez-Mayorga M, Naranjo S, de la Calle-Mustienes E, Maeso I, Beccari L, Aneas I, Vielmas E, et al: **Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112:**7542-7547.

314. Weth O, Renkawitz R: **CTCF function is modulated by neighboring DNA binding factors.** *Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire* 2011, **89:**459-468.

315. Bergstrom R, Savary K, Moren A, Guibert S, Heldin C-H, Ohlsson R, Moustakas A: **Transforming Growth Factor beta Promotes Complexes between Smad Proteins and the CCCTC-binding Factor on the H19 Imprinting Control Region Chromatin.** *Journal of Biological Chemistry* 2010, **285:**19727-19737.

316. Pena-Hernandez R, Marques M, Hilmi K, Zhao T, Saad A, Alaoui-Jamali MA, del Rincon SV, Ashworth T, Roy AL, Emerson BM, Witcher M: **Genome-wide targeting of the epigenetic regulatory protein CTCF to gene promoters by the transcription factor TFII-I.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112:**E677-E686.

317. Liu Z, Scannell DR, Eisen MB, Tjian R: **Control of Embryonic Stem Cell Lineage Commitment by Core Promoter Factor, TAF3.** *Cell* 2011, **146:**720-731.

318. Zhao H, Sifakis EG, Sumida N, Millan-Arino L, Scholz BA, Svensson JP, Chen X, Ronnegren AL, de Lima CDM, Varnoosfaderani FS, et al: **PARP1-and CTCF-Mediated Interactions between Active and Repressed Chromatin at the Lamina Promote Oscillating Transcription.** *Molecular Cell* 2015, **59:**984-997.

319. Heath H, Ribeiro de Almeida C, Sleutels F, Dingjan G, van de Nobelen S, Jonkers I, Ling KW, Gribnau J, Renkawitz R, Grosveld F, et al: **CTCF regulates cell cycle progression of alphabeta T cells in the thymus.** *EMBO J* 2008, **27:**2839-2850.

320. White JK, Gerdin A-K, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ, Podrini C, et al: **Genome-wide Generation and Systematic Phenotyping of Knockout Mice Reveals New Roles for Many Genes.** *Cell* 2013, **154:**452-464.

321. Xu H, Balakrishnan K, Malaterre J, Beasley M, Yan Y, Essers J, Appeldoorn E, Thomaszewski JM, Vazquez M, Verschoor S, et al: **Rad21-Cohesin Haploinsufficiency Impedes DNA Repair and Enhances Gastrointestinal Radiosensitivity in Mice.** *Plos One* 2010, **5**.

322. Moore JM, Rabaia NA, Smith LE, Fagerlie S, Gurley K, Loukinov D, Disteche CM, Collins SJ, Kemp CJ, Lobanenkov VV, Filippova GN: **Loss of Maternal CTCF Is Associated with Peri-Implantation Lethality of Ctcf Null Embryos.** *Plos One* 2012, **7**.

323. Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG: **Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization.** *Cell* 2017, **169:**930-944 e922.

324. Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, et al: **Cohesin Loss Eliminates All Loop Domains.** *Cell* 2017, **171:**305-320 e324.

325. de Almeida CR, Stadhouders R, de Bruijn MJW, Bergen IM, Thongjuea S, Lenhard B, van Ijcken W, Grosveld F, Galjart N, Soler E, Hendriks RW: **The DNA-Binding Protein CTCF Limits Proximal V kappa Recombination and Restricts kappa Enhancer Interactions to the Immunoglobulin kappa Light Chain Locus.** *Immunity* 2011, **35:**501-513.

326. Hirayama T, Tarusawa E, Yoshimura Y, Galjart N, Yagi T: **CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons.** *Cell Rep* 2012, **2:**345-357.

327. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schujiers J, Lee TI, Zhao K, Young RA: **Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes.** *Cell* 2014, **159:**374-387.

328. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IWF, et al: **Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.** *Proc Natl Acad Sci U S A* 2014, **111:**996-1001.

329. Kemp CJ, Moore JM, Moser R, Bernard B, Teater M, Smith LE, Rabaia NA, Gurley KE, Guinney J, Busch SE, et al: **CTCF Haploinsufficiency Destabilizes DNA Methylation and Predisposes to Cancer.** *Cell Reports* 2014, **7:**1020-1029.

330. Ohlsson R, Renkawitz R, Lobanenkov V: **CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease.** *Trends in Genetics* 2001, **17:**520-527.

331. Filippova GN, Lindblom A, Meincke LJ, Klenova EM, Neiman PE, Collins SJ, Doggett NA, Lobanenkov VV: **A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers.** *Genes Chromosomes & Cancer* 1998, **22:**26-36.

332. Aitken SJ, Ibarra-Soria X, Kentepozidou E, Flicek P, Feig C, Marioni JC, Odom DT: **CTCF maintains regulatory homeostasis of cancer pathways.** *Genome Biol* 2018, **19:**106.

333. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N: **IntOGen-mutations identifies cancer drivers across tumor types.** *Nature Methods* 2013, **10:**1081-1082.

334. Rubio-Perez C, Tamborero D, Schroeder MP, Antolin AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, Lopez-Bigas N: **In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities.** *Cancer Cell* 2015, **27:**382-396.

335. Satou Y, Miyazato P, Ishihara K, Yaguchi H, Melamed A, Miura M, Fukuda A, Nosaka K, Watanabe T, Rowan AG, et al: **The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome.** *Proc Natl Acad Sci U S A* 2016, **113:**3054-3059.

336. Melamed A, Yaguchi H, Miura M, Witkover A, Fitzgerald TW, Birney E, Bangham CR: **The human leukemia virus HTLV-1 alters the structure and transcription of host chromatin in cis.** *Elife* 2018, **7**.

337. Liu F, Wu D, Wang X: **Roles of CTCF in conformation and functions of chromosome.** *Semin Cell Dev Biol* 2018.

338. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al: **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.** *Nature Genetics* 2006, **38:**1341-1347.

339. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W: **CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.** *Genes Dev* 2006, **20:**2349-2354.

340. Vostrov AA, Quitschke WW: **The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation.** *J Biol Chem* 1997, **272:**33353-33359.

341. Kitchen NS, Schoenherr CJ: **Sumoylation modulates a domain in CTCF that activates transcription and decondenses chromatin.** *J Cell Biochem* 2010, **111:**665-675.

342. Kim S, Yu NK, Kaang BK: **CTCF as a multifunctional protein in genome regulation and gene expression.** *Exp Mol Med* 2015, **47:**e166.

343. Perez-Juste G, Garcia-Silva S, Aranda A: **An element in the region responsible for premature termination of transcription mediates repression of c-myc gene expression by thyroid hormone in neuroblastoma cells.** *J Biol Chem* 2000, **275:**1307-1314.

344. Lutz M, Burke LJ, Barreto G, Goeman F, Greb H, Arnold R, Schultheiss H, Brehm A, Kouzarides T, Lobanenkov V, Renkawitz R: **Transcriptional repression by the insulator protein CTCF involves histone deacetylases.** *Nucleic Acids Research* 2000, **28:**1707-1713.

345. Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, Goodwin GH: **A NOVEL SEQUENCE-SPECIFIC DNA-BINDING PROTEIN WHICH INTERACTS WITH 3 REGULARLY SPACED DIRECT REPEATS OF THE CCCTC-MOTIF IN THE 5'-FLANKING SEQUENCE OF THE CHICKEN C-MYC GENE.** *Oncogene* 1990, **5:**1743-1753.

346. Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, et al: **Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome.** *Cell* 2012, **149:**1368-1380.

347. Majumder P, Gomez JA, Chadwick BP, Boss JM: **The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions.** *Journal of Experimental Medicine* 2008, **205:**785-798.

348. Majumder P, Boss JM: **CTCF Controls Expression and Chromatin Architecture of the Human Major Histocompatibility Complex Class II Locus.** *Molecular and Cellular Biology* 2010, **30:**4211-4223.

349. Paredes SH, Melgar MF, Sethupathy P: **Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index.** *Bioinformatics* 2013, **29:**1485-1487.

350. Ruiz-Velasco M, Kumar M, Lai MC, Bhat P, Solis-Pinson AB, Reyes A, Kleinsorg S, Noh KM, Gibson TJ, Zaugg JB: **CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals.** *Cell Syst* 2017, **5:**628-637 e626.

351. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.** *Nature* 2011, **479:**74-U99.

352. Chao W, Huynh KD, Spencer RJ, Davidow LS, Lee JT: **CTCF, a candidate trans-acting factor for X-inactivation choice.** *Science* 2002, **295:**345-347.

353. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al: **Cohesin mediates transcriptional insulation by CCCTC-binding factor.** *Nature* 2008, **451:**796-U793.

354. Nasmyth K, Haering CH: **Cohesin: Its Roles and Mechanisms.** In *Annual Review of Genetics. Volume* 43; 2009: 525-558: *Annual Review of Genetics*].

355. Merkenschlager M, Nora EP: **CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation.** *Annu Rev Genomics Hum Genet* 2016, **17:**17-43.

356. Wendt KS, Peters JM: **How cohesin and CTCF cooperate in regulating gene expression.** *Chromosome Res* 2009, **17:**201-214.

357. Unal E, Arbel-Eden A, Sattler U, Shroff R, Lichten M, Haber JE, Koshland D: **DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain.** *Molecular Cell* 2004, **16:**991-1002.

358. Strom L, Lindroos HB, Shirahige K, Sjogren C: **Postreplicative recruitment of cohesin to double-strand breaks is required for DNA repair.** *Molecular Cell* 2004, **16:**1003-1015.

359. Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, Carroll JS, Flicek P, Odom DT: **A CTCF-independent role for cohesin in tissue-specific transcription.** *Genome Res* 2010, **20:**578-588.

360. Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J: **Massively multiplex single-cell Hi-C.** *Nat Methods* 2017, **14:**263-266.

361. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, Schmierer B, Jolma A, Kivioja T, Taipale M, Taipale J: **Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites.** *Cell* 2013, **154:**801-813.

362. Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P: **Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules.** *Genome Res* 2012, **22:**2163-2175.

363. Dorsett D, Merkenschlager M: **Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans.** *Current Opinion in Cell Biology* 2013, **25:**327-333.

364. Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X, et al: **Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation.** *Nature* 2013, **498:**516-+.

365. Oti M, Falck J, Huynen MA, Zhou H: **CTCF-mediated chromatin loops enclose inducible gene regulatory domains.** *Bmc Genomics* 2016, **17**.

366. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, et al: **CTCF-mediated functional chromatin interactome in pluripotent cells.** *Nature Genetics* 2011, **43:**630-U198.

367. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG: **Evolutionarily Conserved Principles Predict 3D Chromatin Organization.** *Molecular Cell* 2017, **67:**837-+.

368. Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X: **CTCF and cohesin regulate chromatin loop stability with distinct dynamics.** *Elife* 2017, **6**.

369. Vian L, Pekowska A, Rao SSP, Kieffer-Kwon KR, Jung S, Baranello L, Huang SC, El Khattabi L, Dose M, Pruett N, et al: **The Energetics and Physiological Impact of Cohesin Extrusion.** *Cell* 2018, **173:**1165-1178 e1120.

370. Hu J, Zhang Y, Zhao L, Frock RL, Du Z, Meyers RM, Meng F-l, Schatz DG, Alt FW: **Chromosomal Loop Domains Direct the Recombination of Antigen Receptor Genes.** *Cell* 2015, **163:**947-959.

371. Davidson IF, Goetz D, Zaczek MP, Molodtsov MI, in't Veld PJH, Weissmann F, Litos G, Cisneros DA, Ocampo-Hafalla M, Ladurner R, et al: **Rapid movement and transcriptional re-localization of human cohesin on DNA.** *Embo Journal* 2016, **35:**2671-2685.

372. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al: **Cohesins functionally associate with CTCF on mammalian chromosome arms.** *Cell* 2008, **132:**422-433.

373. Xiao T, Wallace J, Felsenfeld G: **Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity.** *Molecular and Cellular Biology* 2011, **31:**2174-2183.

374. Remeseiro S, Cuadrado A, Gomez-Lopez G, Pisano DG, Losada A: **A unique role of cohesin-SA1 in gene regulation and development.** *Embo Journal* 2012, **31:**2090-2102.

375. Darwin C: *On the origin of species by means of natural selection.* London,: J. Murray; 1859.

376. Romero IG, Ruvinsky I, Gilad Y: **Comparative studies of gene expression and the evolution of gene regulation.** *Nature Reviews Genetics* 2012, **13:**505-516.

377. Fay JC, Wittkopp PJ: **Evaluating the role of natural selection in the evolution of gene regulation.** *Heredity* 2008, **100:**191-199.

378. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A: **Characterization of evolutionary rates and constraints in three Mammalian genomes.** *Genome Res* 2004, **14:**539-548.

379. Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM: **cis-regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans.** *Cell* 2007, **131:**1179-1189.

380. Tung J, Primus A, Bouley AJ, Severson TF, Alberts SC, Wray GA: **Evolution of a malaria resistance gene in wild primates.** *Nature* 2009, **460:**388-U103.

381. Mc CB: **The origin and behavior of mutable loci in maize.** *Proc Natl Acad Sci U S A* 1950, **36:**344-355.

382. Feschotte C: **Opinion - Transposable elements and the evolution of regulatory networks.** *Nature Reviews Genetics* 2008, **9:**397-405.

383. Bourque G: **Transposable elements in gene regulation and in the evolution of vertebrate genomes.** *Curr Opin Genet Dev* 2009, **19:**607-612.

384. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET: **Evolution of the mammalian transcription factor binding repertoire via transposable elements.** *Genome Res* 2008, **18:**1752-1762.

385. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends in Genetics* 2003, **19:**68-72.

386. Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK: **Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA.** *Bmc Genomics* 2008, **9**.

387. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T: **Widespread contribution of transposable elements to the innovation of gene regulatory networks.** *Genome Res* 2014, **24:**1963-1976.

388. Jacques P-E, Jeyakani J, Bourque G: **The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements.** *Plos Genetics* 2013, **9**.

389. Villar D, Flicek P, Odom DT: **Evolution of transcription factor binding in metazoans - mechanisms and functional implications.** *Nature Reviews Genetics* 2014, **15:**221-233.

390. Taft RJ, Pheasant M, Mattick JS: **The relationship between non-protein-coding DNA and eukaryotic complexity.** *Bioessays* 2007, **29:**288-299.

391. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al: **Comparative genomics reveals insights into avian genome evolution and adaptation.** *Science* 2014, **346:**1311-1320.

392. Ward MC, Wilson MD, Barbosa-Morais NL, Schmidt D, Stark R, Pan Q, Schwalie PC, Menon S, Lukk M, Watt S, et al: **Latent regulatory potential of human-specific repetitive elements.** *Mol Cell* 2013, **49:**262-272.

393. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT: **Species-specific transcription in mice carrying human chromosome 21.** *Science* 2008, **322:**434-438.

394. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.** *Science* 2010, **328:**1036-1040.

395. Wray GA: **The evolutionary significance of cis-regulatory mutations.** *Nature Reviews Genetics* 2007, **8:**206-216.

396. Carroll SB: **Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution.** *Cell* 2008, **134:**25-36.

397. Amoutzias GD, Veron AS, Weiner J, III, Robinson-Rechavi M, Bornberg-Bauer E, Oliver SG, Robertson DL: **One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity.** *Molecular Biology and Evolution* 2007, **24:**827-835.

398. Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, Taipale J: **Conservation of transcription factor binding specificities across 600 million years of bilateria evolution.** *Elife* 2015, **4**.

399. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al: **Principles of regulatory information conservation between mouse and human.** *Nature* 2014, **515:**371-+.

400. Ritter DI, Li Q, Kostka D, Pollard KS, Guo S, Chuang JH: **The Importance of Being Cis: Evolution of Orthologous Fish and Mammalian Enhancer Activity.** *Molecular Biology and Evolution* 2010, **27:**2322-2332.

401. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al: **Enhancer evolution across 20 mammalian species.** *Cell* 2015, **160:**554-566.

402. Galis F, van Dooren TJM, Metz JAJ: **Conservation of the segmented germband stage: robustness or pleiotropy?** *Trends in Genetics* 2002, **18:**504-509.

403. He X, Zhang J: **Toward a molecular understanding of pleiotropy.** *Genetics* 2006, **173:**1885-1891.

404. Papakostas S, Vollestad LA, Bruneaux M, Aykanat T, Vanoverbeke J, Ning M, Primmer CR, Leder EH: **Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions.** *Nature Communications* 2014, **5**.

405. Chesmore KN, Bartlett J, Cheng C, Williams SM: **Complex Patterns of Association between Pleiotropy and Transcription Factor Evolution.** *Genome Biology and Evolution* 2016, **8:**3159-3170.

406. Guillaume F, Otto SP: **Gene Functional Trade-Offs and the Evolution of Pleiotropy.** *Genetics* 2012, **192:**1389-+.

407. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M: **Divergence of transcription factor binding sites across related yeast species.** *Science* 2007, **317:**815-819.

408. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.** *Mol Biol Evol* 2002, **19:**1114-1121.

409. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells.** *Nat Genet* 2010, **42:**631-634.

410. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse.** *Nat Genet* 2007, **39:**730-732.

411. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Stamatoyannopoulos JA, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

412. Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP: **The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb.** *Cell* 2013, **154:**185-196.

413. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen CC, Musselman M, Xie M, West FD, Lewin HA, et al: **Comparative epigenomic annotation of regulatory DNA.** *Cell* 2012, **149:**1381-1392.

414. Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al: **Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution.** *Science* 2014, **346:**1007-1012.

415. Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP: **Evolutionary changes in promoter and enhancer activity during human corticogenesis.** *Science* 2015, **347:**1155-1159.

416. Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, Bickmore WA, Forrest AR, et al: **The frequent evolutionary birth and death of functional promoters in mouse and human.** *Genome Research* 2015, **25:**1546-1557.

417. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al: **The evolution of gene expression levels in mammalian organs.** *Nature* 2011, **478:**343-+.

418. Merkin J, Russell C, Chen P, Burge CB: **Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues.** *Science* 2012, **338:**1593-1599.

419. Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, Eisen MB: **Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression.** *Plos Genetics* 2013, **9**.

420. Emera D, Yin J, Reilly SK, Gockley J, Noonan JP: **Origin and evolution of developmental enhancers in the mammalian neocortex.** *Proceedings of the National Academy of Sciences of the United States of America* 2016, **113:**E2617-E2626.

421. Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A: **Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution.** *Nature Genetics* 2014, **46:**685-692.

422. Carvunis AR, Wang T, Skola D, Yu A, Chen J, Kreisberg JF, Ideker T: **Evidence for a common evolutionary rate in metazoan transcriptional networks.** *Elife* 2015, **4**.

423. Fish A, Chen L, Capra JA: **Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity.** *Genome Biol Evol* 2017, **9:**2615-2625.

424. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P: **Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression.** *Nat Ecol Evol* 2018, **2:**152-163.

425. Ronald J, Akey JM: **The Evolution of Gene Expression QTL in Saccharomyces cerevisiae.** *Plos One* 2007, **2**.

426. Martin D, Pantoja C, Fernandez Minan A, Valdes-Quezada C, Molto E, Matesanz F, Bogdanovic O, de la Calle-Mustienes E, Dominguez O, Taher L, et al: **Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes.** *Nat Struct Mol Biol* 2011, **18:**708-714.

427. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED: **Comparative Epigenomic Analysis of Murine and Human Adipogenesis.** *Cell* 2010, **143:**156-169.

428. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon.** *Nature* 2006, **441:**87-90.

429. Han JS, Szak ST, Boeke JD: **Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.** *Nature* 2004, **429:**268-274.

430. Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang X, Wang L, Saenz-Vash V, Gnirke A, et al: **ZBED6, a Novel Transcription Factor Derived from a Domesticated DNA Transposon Regulates IGF2 Expression and Muscle Growth.** *Plos Biology* 2009, **7**.

431. Lynch VJ, Leclerc RD, May G, Wagner GP: **Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.** *Nat Genet* 2011, **43:**1154-1159.

432. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D: **Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**18613-18618.

433. Mortazavi A, Thompson ECL, Garcia ST, Myers RM, Wold B: **Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire.** *Genome Research* 2006, **16:**1208-1221.

434. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, Flicek P: **Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes.** *Genome Biol* 2013, **14:**R148.

435. Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, Simon I: **Comparative Analysis of DNA Replication Timing Reveals Conserved Large-Scale Chromosomal Architecture.** *Plos Genetics* 2010, **6**.

436. de Souza FS, Franchini LF, Rubinstein M: **Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong?** *Mol Biol Evol* 2013, **30:**1239-1251.

437. Sundaram V, Wang T: **Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells: Re-Visiting the "Gene-Battery" Model.** *Bioessays* 2018, **40**.

438. Zeng L, Pederson SM, Cao D, Qu Z, Hu Z, Adelson DL, Wei C: **Genome-Wide Analysis of the Association of Transposable Elements with Gene Regulation Suggests that Alu Elements Have the Largest Overall Regulatory Impact.** *J Comput Biol* 2018, **25:**551-562.

439. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10:**691-703.

440. Lander ES, Int Human Genome Sequencing C, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

441. Slotkin RK, Martienssen R: **Transposable elements and the epigenetic regulation of the genome.** *Nature Reviews Genetics* 2007, **8:**272-285.

442. Wessler SR: **Transposable elements and the evolution of eukaryotic genomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103:**17600-17601.

443. Feschotte C, Pritham EJ: **DNA transposons and the evolution of eukaryotic genomes.** In *Annual Review of Genetics. Volume* 41; 2007: 331-368: *Annual Review of Genetics*].

444. McLaughlin RN, Jr., Malik HS: **Genetic conflicts: the usual suspects and beyond.** *Journal of Experimental Biology* 2017, **220:**6-17.

445. Biemont C: **A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution.** *Genetics* 2010, **186:**1085-1093.

446. Trizzino M, Kapusta A, Brown CD: **Transposable elements generate regulatory novelty in a tissue-specific fashion.** *BMC Genomics* 2018, **19:**468.

447. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD: **Transposable elements are the primary source of novelty in primate gene regulation.** *Genome Research* 2017, **27:**1623-1633.

448. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Gruetzner F, Bauersachs S, et al: **Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy.** *Cell Reports* 2015, **10:**551-561.

449. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL: **Embryonic stem cell potency fluctuates with endogenous retrovirus activity.** *Nature* 2012, **487:**57-+.

450. Levin HL, Moran JV: **Dynamic interactions between transposable elements and their hosts.** *Nat Rev Genet* 2011, **12:**615-627.

451. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al: **A comparative encyclopedia of DNA elements in the mouse genome.** *Nature* 2014, **515:**355-364.

452. Britten RJ, Davidson EH: **Gene regulation for higher cells: a theory.** *Science* 1969, **165:**349-357.

453. Davidson EH, Britten RJ: **Regulation of gene expression: possible role of repetitive sequences.** *Science* 1979, **204:**1052-1059.

454. Sundaram V, Choudhary MN, Pehrsson E, Xing X, Fiore C, Pandey M, Maricque B, Udawatta M, Ngo D, Chen Y, et al: **Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus.** *Nat Commun* 2017, **8:**14550.

455. Janousek V, Laukaitis CM, Yanchukov A, Karn RC: **The Role of Retrotransposons in Gene Family Expansions in the Human and Mouse Genomes.** *Genome Biol Evol* 2016, **8:**2632-2650.

456. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171:**737-738.

457. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci U S A* 1977, **74:**560-564.

458. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors. 1977.** *Biotechnology* 1992, **24:**104-108.

459. Luckey JA, Drossman H, Kostichka AJ, Mead DA, Dcunha J, Norris TB, Smith LM: **HIGH-SPEED DNA SEQUENCING BY CAPILLARY ELECTROPHORESIS.** *Nucleic Acids Research* 1990, **18:**4417-4421.

460. Swerdlow H, Gesteland R: **CAPILLARY GEL-ELECTROPHORESIS FOR RAPID, HIGH-RESOLUTION DNA SEQUENCING.** *Nucleic Acids Research* 1990, **18:**1415-1419.

461. Jackson DA, Berg P, Symons RH: **BIOCHEMICAL METHOD FOR INSERTING NEW GENETIC INFORMATION INTO DNA OF SIMIAN VIRUS 40 - CIRCULAR SV40 DNA MOLECULES CONTAINING LAMBDA PHAGE GENES AND GALACTOSE OPERON OF ESCHERICHIA-COLI.** *Proceedings of the National Academy of Sciences of the United States of America* 1972, **69:**2904-&.

462. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA: **PRIMER-DIRECTED ENZYMATIC AMPLIFICATION OF DNA WITH A THERMOSTABLE DNA-POLYMERASE.** *Science* 1988, **239:**487-491.

463. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287:**2185-2195.

464. Consortium CeS: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.

465. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.

466. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome.** *Science* 2001, **291:**1304-+.

467. Alekseyev YO, Fazeli R, Yang S, Basran R, Maher T, Miller NS, Remick D: **A Next-Generation Sequencing Primer-How Does It Work and What Can It Do?** *Acad Pathol* 2018, **5:**2374289518766521.

468. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31-46.

469. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends in Genetics* 2008, **24:**133-141.

470. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.

471. Kim HS, Lee H, Shin S-J, Beom S-H, Jung M, Bae S, Lee EY, Park KH, Choi YY, Son T, et al: **Complementary utility of targeted next-generation sequencing and immunohistochemistry panels as a screening platform to select targeted therapy for advanced gastric cancer.** *Oncotarget* 2017, **8:**38389-38398.

472. D'Haene N, Le Mercier M, De Neve N, Blanchard O, Delaunoy M, El Housni H, Dessars B, Heimann P, Remmelink M, Demetter P, et al: **Clinical Validation of Targeted Next Generation Sequencing for Colon and Lung Cancers.** *Plos One* 2015, **10**.

473. Mardis ER: **Next-Generation Sequencing Platforms.** In *Annual Review of Analytical Chemistry, Vol 6. Volume* 6. Edited by Cooks RG, Pemberton JE; 2013: 287-303: *Annual Review of Analytical Chemistry*].

474. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, **26:**1135-1145.

475. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P: **Library construction for next-generation sequencing: Overviews and challenges.** *Biotechniques* 2014, **56:**61-+.

476. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454:**766-770.

477. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.

478. de Wit E, de Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes Dev* 2012, **26:**11-24.

479. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: **ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.** *Curr Protoc Mol Biol* 2015, **109:**21 29 21-29.

480. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: **The next-generation sequencing revolution and its impact on genomics.** *Cell* 2013, **155:**27-38.

481. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T: **Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond.** *Cell Cycle* 2014, **13:**2847-2852.

482. Nakato R, Shirahige K: **Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation.** *Brief Bioinform* 2017, **18:**279-290.

483. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Res* 2012, **22:**1813-1831.

484. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim T-K, He HH, Zieba J, et al: **Systematic evaluation of factors influencing ChIP-seq fidelity.** *Nature Methods* 2012, **9:**609-+.

485. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**14926-14931.

486. Mimura I, Kanki Y, Kodama T, Nangaku M: **Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq.** *Kidney Int* 2014, **85:**31-38.

487. Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, Nagarajan R, Carter AB, Pantanowitz L: **Next-Generation Sequencing Informatics: Challenges and Strategies for Implementation in a Clinical Environment.** *Arch Pathol Lab Med* 2016, **140:**958-975.

488. Andrews S, Lindenbaum P, Howard B, Ewels P: **FastQC High Throughput Sequence QC Report.** 0.11.5 edition: Babraham Bioinformatics; 2011.

489. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nature Reviews Genetics* 2014, **15:**121-132.

490. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol* 2013, **9:**e1003326.

491. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al: **AlleleSeq: analysis of allele-specific expression and binding in a network framework.** *Molecular Systems Biology* 2011, **7**.

492. Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keles S: **Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data.** *Plos Computational Biology* 2011, **7**.

493. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.** *Nat Rev Genet* 2012, **13:**840-852.

494. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.

495. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10**.

496. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.

497. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25:**1966-1967.

498. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26:**873-881.

499. Day DS, Luquette LJ, Park PJ, Kharchenko PV: **Estimating enrichment of repetitive elements from high-throughput sequence data.** *Genome Biology* 2010, **11**.

500. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9:**R137.

501. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *Nat Biotechnol* 2009, **27:**66-75.

502. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nature Biotechnology* 2008, **26:**1293-1300.

503. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nature Methods* 2008, **5:**829-834.

504. Wilbanks EG, Facciotti MT: **Evaluation of Algorithm Performance in ChIP-Seq Peak Detection.** *Plos One* 2010, **5**.

505. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424:**147-151.

506. Zambelli F, Pesole G, Pavesi G: **Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.** *Briefings in Bioinformatics* 2013, **14:**225-237.

507. Zambelli F, Pesole G, Pavesi G: **PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments.** *Nucleic Acids Research* 2013, **41:**W535-W543.

508. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37:**W202-208.

509. Bailey TL, Johnson J, Grant CE, Noble WS: **The MEME Suite.** *Nucleic Acids Res* 2015, **43:**W39-49.

510. Lihu A, Holban S: **A review of ensemble methods for de novo motif discovery in ChIP-Seq data.** *Brief Bioinform* 2015, **16:**964-973.

511. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

512. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11 12 11-34.

513. Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K: **NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors.** *Genome Research* 2013, **23:**1195-1209.

514. Pavesi G: **ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks.** *Adv Biochem Eng Biotechnol* 2017, **160:**1-14.

515. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nat Biotechnol* 2010, **28:**495-501.

516. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities.** *Molecular Cell* 2010, **38:**576-589.

517. Majewski J, Pastinen T: **The study of eQTL variations by RNA-seq: from SNPs to phenotypes.** *Trends in Genetics* 2011, **27:**72-79.

518. Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in cis and trans gene regulation.** *Nature* 2004, **430:**85-88.

519. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou Y-C, Pugh TJ, et al: **Alternative expression analysis by RNA sequencing.** *Nature Methods* 2010, **7:**843-U108.

520. Sun W, Hu Y: **eQTL Mapping Using RNA-seq Data.** *Statistics in Biosciences* 2012, **5:**198-219.

521. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: **MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.** *Genet Epidemiol* 2010, **34:**816-834.

522. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM: **A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.** *Genome Res* 2011, **21:**1728-1737.

523. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.

524. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast.** *Science* 2002, **296:**752-755.

525. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: **Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.** *Nature Genetics* 2003, **35:**57-64.

526. Brem RB, Kruglyak L: **The landscape of genetic complexity across 5,700 gene expression traits in yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**1572-1577.

527. Brem RB, Storey JD, Whittle J, Kruglyak L: **Genetic interactions between polymorphisms that affect gene expression in yeast.** *Nature* 2005, **436:**701-703.

528. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al: **Genetic and epigenetic fine mapping of causal autoimmune disease variants.** *Nature* 2015, **518:**337-343.

529. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG: **Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk.** *Nature Genetics* 2018, **50:**1171-+.

530. Almlof JC, Lundmark P, Lundmark A, Ge B, Maouche S, Goering HHH, Liljedahl U, Enstrom C, Brocheton J, Proust C, et al: **Powerful Identification of Cis-regulatory SNPs in Human Primary Monocytes Using Allele-Specific Gene Expression.** *Plos One* 2012, **7**.

531. Doss S, Schadt EE, Drake TA, Lusis AJ: **Cis-acting expression quantitative trait loci in mice.** *Genome Research* 2005, **15:**681-691.

532. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake TA: **Allele-specific expression and eQTL analysis in mouse adipose tissue.** *Bmc Genomics* 2014, **15**.

533. Lagarrigue S, Martin L, Hormozdiari F, Roux P-F, Pan C, van Nas A, Demeure O, Cantor R, Ghazalpour A, Eskin E, Lusis AJ: **Analysis of Allele-Specific Expression in Mouse Liver by RNA-Seq: A Comparison With Cis-eQTL Identified Using Genetic Linkage.** *Genetics* 2013, **195:**1157-+.

534. Emerson JJ, Li W-H: **The genetic basis of evolutionary change in gene expression levels.** *Philosophical Transactions of the Royal Society B-Biological Sciences* 2010, **365:**2581-2590.

535. Rockman MV, Kruglyak L: **Genetics of global gene expression.** *Nature Reviews Genetics* 2006, **7:**862-872.

536. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al: **Genetics of gene expression and its effect on disease.** *Nature* 2008, **452:**423-U422.

537. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, et al: **Heritability and tissue specificity of expression quantitative trait loci.** *Plos Genetics* 2006, **2:**1625-1633.

538. Fish AE, Capra JA, Bush WS: **Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts?** *American Journal of Human Genetics* 2016, **99:**817-830.

539. Emerson JJ, Hsieh L-C, Sung H-M, Wang T-Y, Huang C-J, Lu HH-S, Lu M-YJ, Wu S-H, Li W-H: **Natural selection on cis and trans regulation in yeasts.** *Genome Research* 2010, **20:**826-836.

540. Schaefke B, Emerson JJ, Wang T-Y, Lu M-YJ, Hsieh L-C, Li W-H: **Inheritance of Gene Expression Level and Selective Constraints on Trans- and Cis-Regulatory Changes in Yeast.** *Molecular Biology and Evolution* 2013, **30:**2121-2133.

541. Tirosh I, Reikhav S, Levy AA, Barkai N: **A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation.** *Science* 2009, **324:**659-662.

542. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ: **Regulatory divergence in Drosophila revealed by mRNA-seq.** *Genome Research* 2010, **20:**816-825.

543. Wittkopp PJ, Haerum BK, Clark AG: **Regulatory changes underlying expression differences within and between Drosophila species.** *Nature Genetics* 2008, **40:**346-350.

544. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES: **Detection of regulatory variation in mouse genes.** *Nature Genetics* 2002, **32:**432-437.

545. Mack KL, Campbell P, Nachman MW: **Gene regulation and speciation in house mice.** *Genome Res* 2016, **26:**451-461.

546. Kempfer R, Pombo A: **Methods for mapping 3D chromosome architecture.** *Nature Reviews Genetics* 2019, **21:**207-226.

547. Speicher MR, Ballard SG, Ward DC: **Karyotyping human chromosomes by combinatorial multi-fluor FISH.** *Nature Genetics* 1996, **12:**368-375.

548. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295:**1306-1311.

549. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).** *Nature Genetics* 2006, **38:**1348-1354.

550. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, et al: **Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements.** *Genome Research* 2006, **16:**1299-1309.

551. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Bin Mohamed Y, Ooi H-S, Tennakoon C, et al: **ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.** *Genome Biology* 2010, **11**.

552. Belaghzal H, Dekker J, Gibcus JH: **Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation.** *Methods* 2017, **123:**56-65.

553. Lajoie BR, Dekker J, Kaplan N: **The Hitchhiker's guide to Hi-C analysis: Practical guidelines.** *Methods* 2015, **72:**65-75.

554. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL: **Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.** *Cell Systems* 2016, **3:**95-98.

555. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S: **HiCUP: Pipeline for mapping and processing Hi-C data.** *F1000Research* 2015, **4**.

556. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biology* 2015, **16**.

557. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies.** *Nat Rev Genet* 2010, **11:**843-854.

558. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al: **Cooperativity and rapid evolution of cobound transcription factors in closely related mammals.** *Cell* 2013, **154:**530-540.

559. Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janousek V, Akanni W, et al: **Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes.** *Genome Res* 2018, **28:**448-459.

560. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, Liu ET: **Evolution of the mammalian transcription factor binding repertoire via transposable elements.** *Genome Research* 2008, **18:**1752-1762.

561. Feschotte C: **Transposable elements and the evolution of regulatory networks.** *Nature Reviews Genetics* 2008, **9:**397-405.

562. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells.** *Nature Genetics* 2010, **42:**631-U111.

563. Kass DH, Kim J, Rao A, Deininger PL: **Evolution of B2 repeats: the muroid explosion.** *Genetica* 1997, **99:**1-13.

564. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al: **Mouse genomic variation and its effect on phenotypes and gene regulation.** *Nature* 2011, **477:**289-294.

565. Doran AG, Wong K, Flint J, Adams DJ, Hunter KW, Keane TM: **Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations.** *Genome Biol* 2016, **17:**167.

566. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Czechanski A, Danecek P, et al: **Multiple laboratory mouse reference genomes define strain specific haplotypes and novel functional loci.** *bioRxiv* 2018.

567. Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT: **ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions.** *Methods* 2009, **48:**240-248.

568. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

569. Hickey G, Paten B, Earl D, Zerbino D, Haussler D: **HAL: a hierarchical format for storing and analyzing multiple genome alignments.** *Bioinformatics* 2013, **29:**1341-1342.

570. **RepeatMasker** [http://repeatmasker.org]

571. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data.** *Bioinformatics* 2016, **32:**2847-2849.

572. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, Simpson G, Solymos P, Stevens MHH, Wagner H: **vegan: Community ecology package. R package version 2.0-7. Online publication.** 2013.

573. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.

574. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al: **Ensembl 2019.** *Nucleic Acids Res* 2019, **47:**D745-D751.

575. Kentepozidou E, Aitken SJ, Feig C, Stefflova K, Ibarra-Soria X, Odom DT, Roller M, Flicek P: **Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains.** *Genome Biol* 2020, **21:**5.

576. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Collins S, Czechanski A, et al: **Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci.** *Nat Genet* 2018, **50:**1574-1583.

577. Ong C-T, Corces VG: **CTCF: an architectural protein bridging genome topology and function.** *Nature Reviews Genetics* 2014, **15:**234-246.

578. Shannon CE: **A mathematical theory of communication.** *The Bell System Technical Journal* 1948, **27:**379-423.

579. van Pesch V, Lanaya H, Renauld JC, Michiels T: **Characterization of the murine alpha interferon gene family.** *J Virol* 2004, **78:**8219-8228.

580. Hardy MP, Owczarek CM, Jermiin LS, Ejdeback M, Hertzog PJ: **Characterization of the type I interferon locus and identification of novel genes.** *Genomics* 2004, **84:**331-345.

581. Xu L, Yang L, Liu W: **Distinct evolution process among type I interferon in mammals.** *Protein Cell* 2013, **4:**383-392.

582. Mudge JM, Harrow J: **Creating reference gene annotation for the mouse C57BL6/J genome assembly.** *Mamm Genome* 2015, **26:**366-378.

583. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.

584. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.

585. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

586. The Gene Ontology C: **Expansion of the Gene Ontology knowledgebase and resources.** *Nucleic Acids Res* 2017, **45:**D331-D338.

587. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al: **Analyses of pig genomes provide insight into porcine demography and evolution.** *Nature* 2012, **491:**393-398.

588. Roman AC, Gonzalez-Rico FJ, Molto E, Hernando H, Neto A, Vicente-Garcia C, Ballestar E, Gomez-Skarmeta JL, Vavrova-Anderson J, White RJ, et al: **Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch.** *Genome Res* 2011, **21:**422-432.

589. Yokoyama KD, Zhang Y, Ma J: **Tracing the evolution of lineage-specific transcription factor binding sites in a birth-death framework.** *PLoS Comput Biol* 2014, **10:**e1003771.

590. Chen H, Tian Y, Shu W, Bo X, Wang S: **Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome.** *PLoS One* 2012, **7:**e41374.

591. Glinsky GV: **Contribution of transposable elements and distal enhancers to evolution of human-specific features of interphase chromatin architecture in embryonic stem cells.** *Chromosome Res* 2018, **26:**61-84.

592. Janousek V, Karn RC, Laukaitis CM: **The role of retrotransposons in gene family expansions: insights from the mouse Abp gene family.** *BMC Evol Biol* 2013, **13:**107.

593. Oritani K, Medina KL, Tomiyama Y, Ishikawa J, Okajima Y, Ogawa M, Yokota T, Aoyama K, Takahashi I, Kincade PW, Matsuzawa Y: **Limitin: An interferon-like cytokine that preferentially influences B-lymphocyte precursors.** *Nat Med* 2000, **6:**659-666.

594. Lemos B, Araripe LO, Fontanillas P, Hartl DL: **Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**14471-14476.

595. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK: **The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans.** *Nature Genetics* 2005, **37:**544-548.

596. Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in cis and trans gene regulation.** *Nature* 2004, **430:**85-88.

597. Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL: **Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila.** *Genetics* 2005, **171:**1813-1822.

598. Graze RM, McIntyre LM, Main BJ, Wayne ML, Nuzhdin SV: **Regulatory Divergence in Drosophila melanogaster and D. simulans, a Genomewide Analysis of Allele-Specific Expression.** *Genetics* 2009, **183:**547-561.

599. Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ: **Tempo and mode of regulatory evolution in Drosophila.** *Genome Research* 2014, **24:**797-808.

600. Shen SQ, Turro E, Corbo JC: **Hybrid Mice Reveal Parent-of-Origin and Cis- and Trans-Regulatory Effects in the Retina.** *Plos One* 2014, **9**.

601. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al: **Variation in Transcription Factor Binding Among Humans.** *Science* 2010, **328:**232-235.

602. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA: **Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo.** *Nature Genetics* 2015, **47:**1393-+.

603. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A, et al: **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans.** *Cell* 2015, **162:**1039-1050.

604. Levo M, Zalckvar E, Sharon E, Machado ACD, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R, Segal E: **Unraveling determinants of transcription factor binding outside the core binding site.** *Genome Research* 2015, **25:**1018-1029.

605. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al: **The human transcriptome across tissues and individuals.** *Science* 2015, **348:**660-665.

606. Uhlen M, Fagerberg L, Hallstroem BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjoestedt E, Asplund A, et al: **Tissue-based map of the human proteome.** *Science* 2015, **347**.

607. McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al: **Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans.** *Science* 2010, **328:**235-239.

608. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al: **DNase I sensitivity QTLs are a major determinant of human expression variation.** *Nature* 2012, **482:**390-394.

609. Ding Z, Ni Y, Timmer SW, Lee B-K, Battenhouse A, Louzada S, Yang F, Dunham I, Crawford GE, Lieb JD, et al: **Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association.** *Plos Genetics* 2014, **10**.

610. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.

611. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biology* 2010, **11**.

612. Wittkopp PJ, Kalay G: **Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.** *Nature Reviews Genetics* 2012, **13:**59-69.

613. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A, et al: **Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions.** *Cell* 2015, **162:**1051-1065.

614. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA: **Circuitry and dynamics of human transcription factor regulatory networks.** *Cell* 2012, **150:**1274-1286.

615. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al: **Effects of sequence variation on differential allelic transcription factor occupancy and gene expression.** *Genome Res* 2012, **22:**860-869.

616. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, et al: **Mapping cis- and trans-regulatory effects across multiple tissues in twins.** *Nat Genet* 2012, **44:**1084-1089.

617. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK: **Effect of natural genetic variation on enhancer selection and function.** *Nature* 2013, **503:**487-+.

618. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al: **Conservation of trans-acting circuitry during mammalian regulatory evolution.** *Nature* 2014, **515:**365-370.

619. Kruglyak L, Stern DL: **Evolution. An embarrassment of switches.** *Science* 2007, **317:**758-759.

620. Stone JR, Wray GA: **Rapid evolution of cis-regulatory sequences via local point mutations.** *Molecular Biology and Evolution* 2001, **18:**1764-1770.

621. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT: **Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.** *Cell* 2012, **148:**335-348.

622. Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S: **Cohesin-mediated interactions organize chromosomal domain architecture.** *EMBO Journal* 2013, **32:**3119-3129.

623. Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, van de Corput MPC, van de Werken HJG, Knoch TA, van Ijcken WFJ, et al: **Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.** *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111:**996-1001.

624. Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, et al: **3D Chromosome Regulatory Landscape of Human Pluripotent Cells.** *Cell Stem Cell* 2016, **18:**262-275.

625. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485:**381-385.

626. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B: **Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation.** *Nat Commun* 2017, **8:**441.

627. Ren G, Jin W, Cui K, Rodrigez J, Hu G, Zhang Z, Larson DR, Zhao K: **CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression.** *Molecular Cell* 2017, **67:**1049-+.

628. Busslinger GA, Stocsits RR, Van Der Lelij P, Axelsson E, Tedeschi A, Galjart N, Peters JM: **Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl.** *Nature* 2017, **544:**503-507.

629. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al: **GENCODE reference annotation for the human and mouse genomes.** *Nucleic Acids Res* 2019, **47:**D766-D773.

630. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38:**576-589.

631. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P: **PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci.** *BMC Bioinformatics* 2010, **11:**415.

632. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based Analysis of ChIP-Seq (MACS).** *Genome Biology* 2008, **9**.

633. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance genomic feature operations.** *Bioinformatics* 2012, **28:**1919-1920.

634. Ziebarth JD, Bhattacharya A, Cui Y: **CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization.** *Nucleic Acids Res* 2013, **41:**D188-194.

635. Bao L, Zhou M, Cui Y: **CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators.** *Nucleic Acids Res* 2008, **36:**D83-87.

636. Ambrosini G, Groux R, Bucher P: **PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix.** *Bioinformatics* 2018, **34:**2483-2484.

637. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639-1645.

638. Fudenberg G, Pollard KS: **Chromatin features constrain structural variation across evolutionary timescales.** *Proc Natl Acad Sci U S A* 2019, **116:**2175-2180.

639. Holwerda SJ, de Laat W: **CTCF: the protein, the binding partners, the binding sites and their chromatin loops.** *Philos Trans R Soc Lond B Biol Sci* 2013, **368:**20120369.

640. Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Rudan MV, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S: **Cohesin-mediated interactions organize chromosomal domain architecture.** *Embo Journal* 2013, **32:**3119-3129.

641. Stedman W, Kang H, Lin S, Kissil JL, Bartolomei MS, Lieberman PM: **Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators.** *Embo Journal* 2008, **27:**654-666.

642. Ziebarth JD, Bhattacharya A, Cui Y: **CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization.** *Nucleic Acids Research* 2012, **41:**D188-D194.

643. Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janousek V, Akanni W, et al: **Repeat associated mechanisms of genome evolution and function revealed by the &ITMus caroli&IT and &ITMus pahari&IT genomes.** *Genome Research* 2018, **28:**448-459.

644. Diehl AG, Ouyang N, Boyle AP: **Transposable elements strongly contribute to cell-specific and species-specific looping diversity in mammalian genomes.** *bioRxiv* 2019**:**679217.

645. Gothe HJ, Bouwman BAM, Gusmao EG, Piccinno R, Petrosino G, Sayols S, Drechsel O, Minneker V, Josipovic N, Mizi A, et al: **Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations.** *Molecular Cell* 2019, **75:**267-283.e212.

646. Oomen ME, Hansen AS, Liu Y, Darzacq X, Dekker J: **CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning.** *Genome Res* 2019, **29:**236-249.

647. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al: **CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.** *Cell* 2015, **163:**1611-1627.

648. Barrington C, Georgopoulou D, Pezic D, Varsally W, Herrero J, Hadjur S: **Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology.** *Nat Commun* 2019, **10:**2908.

649. Hansen AS, Amitai A, Cattoglio C, Tjian R, Darzacq X: **Guided nuclear exploration increases CTCF target search efficiency.** *Nat Chem Biol* 2019.

650. Hansen AS, Hsieh THS, Cattoglio C, Pustova I, Saldaña-Meyer R, Reinberg D, Darzacq X, Tjian R: **Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF.** *Molecular Cell* 2019, **76:**395-411.e313.

651. Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG: **Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization.** *Cell* 2017, **169:**930-+.

652. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T: **Co-opted transposons help perpetuate conserved higher-order chromosomal structures.** *Genome Biol* 2020, **21:**16.

653. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, et al: **Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer.** *Nature Genetics* 2016, **48:**500-+.

654. Monahan K, Rudnick ND, Kehayova PD, Pauli F, Newberry KM, Myers RM, Maniatis T: **Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-alpha gene expression.** *Proc Natl Acad Sci U S A* 2012, **109:**9125-9130.

655. Saldaña-Meyer R, González-Buendía E, Guerrero G, Narendra V, Bonasio R, Recillas-Targa F, Reinberg D: **CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53.** *Genes and Development* 2014, **28:**723-734.

656. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA: **Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes.** *Cell* 2014, **159:**374-387.

657. Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, Fisher AG, et al: **Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments.** *Genome Research* 2013, **23:**2066-2077.

658. Matthews BJ, Waxman DJ: **Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver.** *Elife* 2018, **7**.

659. Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W: **Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features.** *Biorxiv* 2017.

660. Li W, Notani D, Rosenfeld MG: **Enhancers as non-coding RNA transcription units: recent insights and future perspectives.** *Nature Reviews Genetics* 2016, **17:**207-223.

661. Symmons O, Pan L, Remeseiro S, Aktas T, Klein F, Huber W, Spitz F: **The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances.** *Developmental Cell* 2016, **39:**529-543.

662. Spivakov M: **Spurious transcription factor binding: Non-functional or genetically redundant?** *Bioessays* 2014, **36:**798-806.

663. Rada-Iglesias A, Grosveld FG, Papantonis A: **Forces driving the three-dimensional folding of eukaryotic genomes.** *Mol Syst Biol* 2018, **14:**e8214.

664. Behera V, Evans P, Face CJ, Hamagami N, Sankaranarayanan L, Keller CA, Giardine B, Tan K, Hardison RC, Shi J, Blobel GA: **Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility.** *Nature Communications* 2018, **9**.

665. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, Xu J, Yuan G-C: **Dissecting super-enhancer hierarchy based on chromatin interactions.** *Nature Communications* 2018, **9**.

666. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.

667. Gregor A, Oti M, Kouwenhoven EN, Hoyer J, Sticht H, Ekici AB, Kjaergaard S, Rauch A, Stunnenberg HG, Uebe S, et al: **De novo mutations in the genome organizer CTCF cause intellectual disability.** *Am J Hum Genet* 2013, **93:**124-131.

668. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T, et al: **CTCF/cohesin-binding sites are frequently mutated in cancer.** *Nature Genetics* 2015, **47:**818-+.

669. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A, et al: **Parental origin of sequence variants associated with complex diseases.** *Nature* 2009, **462:**868-874.

670. Zhang R, Wang Y, Yang Y, Zhang Y, Ma J: **Predicting CTCF-mediated chromatin loops using CTCF-MP.** *Bioinformatics* 2018, **34:**i133-i141.

671. Brind'Amour J, Liu S, Hudson M, Chen C, Karimi MM, Lorincz MC: **An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations.** *Nat Commun* 2015, **6:**6033.

672. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W, O'Shaughnessy-Kirwan A, et al: **3D structures of individual mammalian genomes studied by single-cell Hi-C.** *Nature* 2017, **544:**59-+.